Cognitive models of decision making

Mehlhorn, Sabine Katja

Cognitive Models of Decision Making

– why precision matters –

Sabine Katja Mehlhorn

Printed by Gildeprint Drukkerijen,
Enschede, the Netherlands.

ISBN printed version: 978-90-367-5322-7
ISBN digital version: 978-90-367-5323-4

RIJKSUNIVERSITEIT GRONINGEN

Cognitive Models of Decision Making

– why precision matters –

Proefschrift

ter verkrijging van het doctoraat in de
Wiskunde en Natuurwetenschappen
aan de Rijksuniversiteit Groningen
op gezag van de
Rector Magnificus, dr. E. Sterken,
in het openbaar te verdedigen op
vrijdag 20 april 2012
om 12:45 uur

door

Sabine Katja Mehlhorn

geboren op 10 juli 1982
te Karl-Marx-Stadt, Duitse Democratische Republiek

Promotor:	Prof. dr. N.A. Taatgen
Beoordelingscommissie:	Prof. dr. C. González
	Prof. dr. J.F. Krems
	Prof. dr. R. Verbrugge

Contents

Introduction

In which I give an overview of the topic and introduce the methods.

Chapter 1

Introduction

Introduction

It was at the beginning of my graduate time, when, during a guest lecture at the University of Chemnitz, Frank Ritter talked about his favorite scientific articles. One of them was Allen Newell's 20 questions paper (1973). Newell had argued that psychology focuses too much on isolated, experimental phenomena and simplifying dichotomies, rather than working towards a precise and unified theory of cognition. If I would have read this paper in more detail then, and truly understood what Newell meant, working on my dissertation might have gone more smoothly. But I did not do that. Rather, I began working in "good psychological tradition". I had studied my theories, I knew how to set up experiments and do an ANOVA, and I thought that was sufficient to investigate cognition.

The starting point of my dissertation was the idea that automatic memory processes are an important aspect underlying decision making. Specifically I was interested in the role of memory activation in diagnostic reasoning. Diagnostic reasoning is the reasoning from observed data to explanations and involves the generation and evaluation of hypotheses that represent potential explanations. I wanted to know why, when confronted with a number of medical symptoms, possible diagnoses seem to pop up almost effortlessly in a physician's head. And, why, when being in a certain context, one cannot help but interpret new information in the light of this context. My idea was that these phenomena were largely due to automatic memory processes, which make information that is associated to the current context (e.g., observed medical symptoms) available in memory. Such available information could then be subjected to more deliberate reasoning processes as they had been classically discussed in the reasoning literature. While the idea of automatic activation processes regulating the availability of memory contents was not new, direct experimental evidence for such memory processes in diagnostic reasoning was sparse.

With the goal to present such evidence, we set out to conduct a series of experiments (Baumann, Mehlhorn, & Bocklisch, 2007; Mehlhorn, Baumann, & Bocklisch, 2008). In these experiments, we used a probe reaction task to track the availability of different diagnostic hypotheses in memory, while participants had to generate diagnoses for sequentially presented medical symptoms. The probe reaction task was based on the idea of lexical decision tasks, where participants respond faster to a probe that is more highly activated in memory than to a probe of lower activation (e.g., Meyer & Schvaneveldt, 1971). If observations indeed activate associated explanations in memory, then, when presented with symptoms like fever, nausea, and headache, a participant should react faster to the probe "influenza", than to a probe that is less related to these symptoms (e.g., "pregnancy"), or to a neutral probe (e.g., "house"). To avoid the possible influence of previous experience on memory activation, in the experiments we used artificial medical knowledge, which consisted of medical symptoms that were caused by hypothetical chemicals. Chemicals were named with single letters, which allowed us to use letters in the probe reaction task, thereby preventing potential problems associated with the use of complete words (e.g., individual differences in reading speed and word frequency effects).

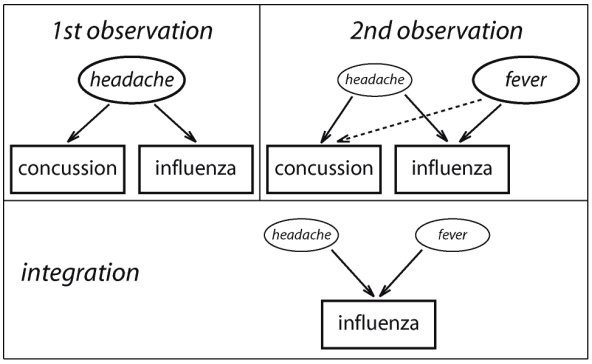


Figure 1.1 Box and arrow model of memory processes assumed to underlie diagnostic reasoning as reported in Baumann et al. (2007).

Overall, the results seemed to support our theoretical considerations. For example, diagnostic hypotheses that were compatible to all observed symptoms caused the fastest probe reactions, suggesting that these compatible hypotheses were indeed more easily available in memory than their alternatives. Also, with an increasing amount of observed symptoms, reaction times to compatible hypotheses decreased faster than reaction times to other hypotheses. This suggested that the availability of hypotheses indeed might be a function of their association to observed symptoms. However, when trying to understand the results in detail, we quickly ran into open questions. Could it have been that the results were actually not caused by memory activation, but were merely byproducts of deliberate reasoning processes? And, assuming that it was indeed memory activation that caused our results, how would the underlying activation processes look precisely? What we needed was a detailed model of the assumed memory processes that would make precise predictions.

The first "model" that we generated consisted of several boxes and arrows (see Figure 1.1 and Baumann et al., 2007). The boxes represented medical symptoms (e.g., headache) and their potential explanations (e.g., influenza). The arrows represented the associations between symptoms and explanations that could be positive (solid lines) or negative (dashed lines). This model was useful to illustrate the processes that we assumed to cause the activation of diagnostic hypotheses. For example, we proposed that "After the integration phase the influenza explanation as a still relevant explanation should be strengthened as it receives activation both from the symptom fever and the symptom headache [...]" (Baumann et al., 2007, p. 804). However, the problem with this model was a lack of precision. For example, why exactly was there a positive association between headache and influenza and how strong was it precisely? As Allen Newell (1973) put it, in such a model "Too much is left unspecific and unconstrained." (p. 301).

The lack of theoretical precision, which we faced when trying to understand our data, is inherent to verbal theories of cognition (and their associated box and arrow models). A solution for that problem has been proposed by Allen Newell (1990) and many others. It is the use of computational cognitive models. These models should be specific and constrained enough to provide quantitative predictions that can be tested by comparing them to human data. After the initial difficulties described above, I moved on to using such models. In the remainder of the introduction I give a brief introduction of the two modeling approaches that I used in my dissertation and present an overview of the chapters in this thesis.

A Connectionist Approach: ECHO

In his *theory of explanatory coherence*, Thagard (1989a, 1989b, 2000) proposes that a set of propositions (e.g., observations and their potential explanations in memory) can be evaluated by automatic activation processes, purely on the basis of their coherence. In the connectionist constraint-satisfaction implementation of this theory, ECHO (e.g., Thagard, 1989a), propositions are represented by a network of interconnected nodes. The connections between the nodes represent the relations (constraints) between the respective propositions. Depending on these connections, when the network is integrated, activation or inhibition is spread between the nodes. After the network has been integrated, the strength of a proposition is indicated by the numerical activation of its node, which depends on its coherence to the other nodes in the network. Applying Thagard's theory to diagnostic reasoning predicts that those explanations that are strongly associated with the observed data are most strongly available in memory (because they receive a large amount of activation) and that less strongly associated explanations have a lower availability (because they receive less activation and potentially also inhibition).

As we will show in Chapter 3, such a connectionist account increases the precision compared to mere verbal predictions. It requires a detailed specification of the assumed memory processes (e.g., how strong is the connection between observation x and explanation y?) and it predicts precise numerical activation values that can be compared to behavioral data. However, this account has also some major limitations (see e.g., Fodor & Pylyshyn, 1988, for an overview). Maybe most importantly, it does not represent a fully functioning cognitive system. While presenting a precise account of activation dynamics within an assumed network, it remains mute about the interplay of these dynamics with, for example, perceptual, decisional, intentional, and motor processes, which might play an important role in human reasoners. Another problematic point is the interpretability of its results. The model predicts precise activation values, which can be plotted against and correlated with behavioral data. But what exactly do these values mean and how, precisely, do they correspond with behavioral data?

An Architectural Approach: ACT-R

An approach that not only endeavors precision, but also comprehensiveness in terms of understanding how the brain "achieves the function of the mind" (Anderson, 2007, p. 7) is the use of cognitive architectures[1]. The term "cognitive architecture" was, maybe most prominently, described by Allen Newell (1990) as a way towards the "ultimate goal" of a unified theory of human cognition. The idea is that the architecture is both a psychological theory, as well as a platform for constructing computational models, that allows for investigating different phenomena within one framework. This idea has been developed since then, resulting in various architectures, like for example EPIC (Meyer & Kieras, 1997), Soar (A. Newell, 1990), and ACT-R (Anderson et al., 2004).

The cognitive architecture I used in my dissertation is ACT-R, because it puts a strong emphasis on processes underlying memory activation. It has received empirical support and validation from a large number of studies in a variety of research areas, ranging from list memory (Anderson, Bothell, Lebiere, & Matessa, 1998) to car driving (Salvucci, 2006). ACT-R allows for modeling of the complete task as solved by the participant. Thereby, without requiring additional assumptions about how the model maps on the experiment, it produces results that are directly comparable to human data. This is possible because the underlying theory makes precise predictions not only about the probability and latency of retrieving facts from memory, but also about the time needed to perceive stimuli and give responses.

In ACT-R, cognition is described by a number of independent modules. Each of the modules represents a different cognitive resource and is associated with specific brain regions. For example, a visual module allows ACT-R to perceive visual stimuli and a motor module allows for motor actions like pressing a key. Most important for the work presented in this thesis are three of the central cognitive modules: the imaginal module, the declarative module, and the procedural module.

The imaginal module holds information necessary to perform the current task and is thereby comparable to the focus of attention in working memory (e.g., Borst, Taatgen, & van Rijn, 2010). In a diagnostic reasoning task, the imaginal module might, for example, hold observed medical symptoms, which determine the present usefulness of potential explanations. In Chapter 2 we investigate how such observed symptoms can affect the availability of explanations in long-term memory.

The declarative module allows for the storage in and retrieval of facts from declarative memory and thereby represents ACT-R's account of long-term memory. In a diagnostic reasoning task, such facts could, for example, be possible diagnoses. Availability of the facts is determined by their activation (see Chapters 2, 4, and 5 for a detailed description of the underlying equations). Basically, the activation of a fact

[1] *In the literature, the term cognitive architecture has also been used for connectionist models (e.g., Kintsch, 1998). In this thesis we use the term cognitive architecture exclusively for what Fodor and Pylyshyn (1988) referred to as "Classical architectures", that is, architectures that are committed to a symbol-level of representation and thereby aspire "paying attention to three things: the brain, the mind (functional cognition), and the architectural abstractions that link them" (Anderson, 2007, p. 8). However, as Fodor and Pylyshyn (1988) point out, connectionism might provide "an account of the neural [...] structures in which Classical cognitive architecture is implemented" (p. 3). In fact, Lebiere and Anderson (1993) successfully created such a connectionist implementation of an early version of ACT-R.*

represents the likelihood that it will be needed in the near future and depends on two factors: its past and present usefulness. In Chapter 4 we explore the respective contribution of these two factors for the availability of hypotheses in diagnostic reasoning.

The procedural module allows for communication between the other modules. It contains production rules, which can recognize patterns of information in the modules' so-called buffers, and react to these patterns by sending requests to the modules. A production rule might, for example, recognize that a visually presented symptom was encoded in the visual buffer and react by requesting the name of this symptom from declarative memory. Production rules implement strategies that the reasoner might use in a certain situation. For example, after retrieving an explanation for observed medical symptoms from memory, one strategy might be to simply give that explanation as diagnosis, whereas another strategy would be to deliberately test the explanation against potential alternatives. In Chapter 5 we use this module to implement different decision making strategies and test how well these strategies predict behavioral data.

Overview

In this thesis I will show how we used the approaches outlined above to implement and test precise models of decision making.

In Chapter 2, we introduce our idea of how memory activation affects the availability of explanations. We present several ACT-R models that all share the assumption that observations stored in working memory can activate associated explanations in long-term memory. The models differ in their assumptions about how sequentially observed symptoms affect the activation of associated explanations over time. Using ACT-R allows us for testing these assumptions within a well-established and elaborate theory of human memory. It also allows for investigating the interaction of the assumed memory processes with other potentially task-relevant factors. The results of the models are compared to human data from two behavioral experiments in which we used the probe reaction task mentioned above to track the availability of different explanations during a sequential diagnostic reasoning task.

In Chapter 3, we explore different methods of modeling sequential information integration with connectionist constraint satisfaction models, based on Thagard's ECHO. Just like the ACT-R models presented in Chapter 2, the models share the basic assumption that observations can activate associated explanations, but they differ in how sequentially observed medical symptoms affect the activation of explanations over time. The models are evaluated on the probe reaction data from the same experiments as presented in Chapter 2.

In Chapter 4, we investigate how an explanation's present usefulness, as reflected by the observed symptoms, interacts with its past usefulness, as reflected by the recency and frequency of previous encounters with the explanation. We thereby test whether the memory mechanisms as proposed by the ACT-R theory can explain why, out of all possible hypotheses, reasoners tend to generate those hypotheses from memory

that have a high a priori probability and a high usefulness in the current context. Model predictions are compared to behavioral data from an experiment in which we manipulated both memory components independently, by means of a secondary task that had to be solved next to a primary diagnostic reasoning task.

In Chapter 5, we move on to a slightly different domain of decision making. Whereas in Chapters 2 to 4 we investigate how automatic activation process affect the availability of information in memory as a function of the past and present environment, in Chapter 5 we investigate how reasoners use information from memory, given its availability. More specifically, we focus on a debate that has evolved over the last decade in the decision-making literature and is centered on the question whether decisions can better be described by simple non-compensatory heuristics or by more complex compensatory decision making strategies. In Chapter 5 we show how the precision and comprehensiveness provided by a cognitive architecture can be used to get beyond the simple dichotomy of non-compensatory versus compensatory decision strategies. We use ACT-R to implement various strategies that have been discussed in the literature and compare the resulting quantitative predictions to behavioral data from two previously published experiments (Pachur, Bröder, & Marewski, 2008).

The Availability of Explanations in Memory for Diagnostic Reasoning

In which I use behavioral experiments and ACT-R models to test the idea that observations activate associated explanations in memory.

An earlier version of this chapter was published as:

Mehlhorn, K., Taatgen, N.A., Lebiere, C., Krems, J.F. (2011). Memory activation and the availability of explanations in sequential diagnostic reasoning. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 37*, 1391-1411.

Chapter 2

The Availability of Explanations in Memory for Diagnostic Reasoning

Chapter 2

Abstract

In the field of diagnostic reasoning, it has been argued that memory activation can provide the reasoner with a subset of possible explanations from memory that are highly adaptive for the task at hand. However, few studies have experimentally tested this assumption. Even less empirical and theoretical work has investigated how newly incoming observations affect the availability of explanations in memory over time. In this chapter we present the results of two experiments in which we address these questions. While participants diagnosed sequentially presented medical symptoms, the availability of potential explanations in memory was measured with an implicit probe reaction-time task. The results of the experiments were used to test four quantitative cognitive models. The models share the general assumption that observations can activate and inhibit explanations in memory. They vary with respect to how newly incoming observations affect the availability of explanations over time. The data of both experiments were predicted best by a model in which all observations in working memory have the same potential to activate explanations from long-term memory and in which these observations do not decay. The results illustrate the power of memory activation processes and show where additional deliberate reasoning strategies might come into play.

Introduction

A basic goal of human cognition is to explain and understand the events happening in the world. Whether it is in scientific discovery, medical diagnosis, software debugging, or social attribution, people try to find explanations based on what they observe. The kind of reasoning underlying this task is often called abductive (Josephson & Josephson, 1996) or diagnostic reasoning (Kim & Keil, 2003) and it is described as highly complex. First, complexity arises from the large number of potential observations that can each have a large number of potential explanations. Take for example a physician who is confronted with a patient's symptoms. Each of the symptoms has a number of possible alternative explanations and only the combination of symptoms allows for selecting a diagnosis. The task is further complicated by the fact that information often does not become available all at once, but only over time. Even if given all at once, observations might be perceived and understood only over time due to limited cognitive capacities. Thus, the ability to integrate newly incoming information over the course of the diagnosis process is important. A related factor is uncertainty. The physician can never be sure if all symptoms necessary to find the correct diagnosis were observed and whether all observed symptoms were caused by the current disease. Despite all these constraints, people often generate explanations with high speed and accuracy (T. R. Johnson & Krems, 2001).

Theories trying to understand diagnostic reasoning consistently make the distinction between, on the one hand, the generation of a potential set of explanations or hypotheses and, on the other hand, the evaluation of these explanations or hypotheses against potential alternatives. Often the evaluation of hypotheses is assumed to be performed in a second, deliberate reasoning stage after a first stage in which potential hypotheses are generated from memory (e.g., Evans, 2006; Kintsch, 1998; Thomas, Dougherty, Sprenger, & Harbison, 2008; Wang, Johnson, & Zhang, 2006a). For the deliberate stage of hypothesis evaluation, a number of strategies that allow reasoners to deal with the complexity of the task have been investigated (cf. T. R. Johnson & Krems, 2001). However, a key aspect of diagnostic reasoning is that observations can be associated with a large number of possible explanations in memory (in fact, the number of potential explanations has been shown to be computationally intractable; Bylander, Allemang, Tanner, & Josephson, 1991). Generating and deliberately evaluating the complete set of explanations is therefore often impossible due to constraints set by cognitive capacity and time available for diagnosis (Dougherty & Hunter, 2003a, 2003b). Consequently, already during the generation of explanations from memory a selection amongst potential alternative hypotheses has to be made (Dougherty, Thomas, & Lange, 2010; Thomas et al., 2008).

The goal of this chapter is to more closely investigate how memory activation processes can provide the reasoner with such an adaptive selection. Specifically, we want to test how memory activation can help the reasoner to select amongst a large number of potential explanations and how this selection is affected by newly observed pieces of information over time. In the remainder of the introduction we first give

a short overview of empirical findings on hypothesis generation and then we take a closer look at the theoretical background.

Empirical Findings on the Generation of Explanations

Thomas et al. (2008) stated, "Although the evaluation of prespecified hypotheses has been the subject of research for many years, relatively little research has been concerned with the initial generation of the to-be-judged hypotheses." (p. 158; see also: Weber, Böckenholt, Hilton, & Wallace, 1993). Existing empirical findings concerning hypothesis generation consistently show that reasoners generate only a subset of up to four possible hypotheses from memory (Barrows, Norman, Neufeld, & Feightner, 1982; Dougherty, Gettys, & Thomas, 1997; Dougherty & Hunter, 2003a; Elstein, Shulman, & Sprafka, 1978; Joseph & Patel, 1990; Mehle, 1982; Weber et al., 1993). Whereas this small number of generated hypotheses seems to contradict the large number of potential hypotheses, research has shown that the selection of hypotheses into the generated subset is highly adaptive. Out of all potential hypotheses, reasoners generate those hypotheses that have a high likelihood of being relevant as explanations in the current situation. Specifically, those hypotheses seem to be generated that (a) have a high a priori probability based on previous experiences (Dougherty et al., 1997; Dougherty & Hunter, 2003a; Gettys, Pliske, Manning, & Casey, 1987; Sprenger & Dougherty, 2006; Weber et al., 1993) and that (b) are most likely in the context of the current observations (Weber et al., 1993).

Although the studies mentioned above say something about the outcome of the hypothesis generation process, they say little about the cognitive processes that yield this outcome (exceptions are Dougherty & Hunter, 2003b, and Dougherty & Sprenger, 2006, who showed that participants tended to generate those hypotheses that were most "active" as defined by a strength manipulation in the learning phase). To test if memory activation can indeed help the reasoner to select explanations from memory, the availability of explanations has to be assessed as a function of the observed information. In previous experiments, the availability of explanations has been estimated using explicit measures. For example, Wang, Johnson and Zhang (2006b) asked their participants for explicit belief ratings after serially presented observations, and Dougherty and Hunter (2003b) asked their participants for probability judgments of different explanations. However, such explicit measures have two major drawbacks. First, explicitly asking participants during the course of the task might influence the outcome of the task itself (cf. Hogarth & Einhorn, 1992). Second, although there have been efforts at clarifying this issue (Drewitz & Thüring, 2009; Thomas et al., 2008), it is not clear how the implicit concept of availability in memory translates into explicit concepts like ratings and judgments. Furthermore, to investigate how the availability of explanations is affected by newly incoming observations, availability should be tracked over time. With few exceptions (Baumann, Krems, & Ritter, 2010; Sprenger, 2007; Wang et al., 2006b) this issue has received little attention in previous studies.

Methods used in diagnostic reasoning research range from protocol analysis of physicians explaining a patient's pathophysiology (Arocha, Wang, & Patel, 2005) to

simple laboratory experiments where only a few pieces of evidence and a few alternative hypotheses need to be considered (e.g., Wang et al., 2006b). Whereas the first method allows for high external validity of aspects like task complexity, the second method allows for high control of aspects like previous knowledge. For analyzing the subtle effects of memory activation it is essential to have an optimal trade-off between both.

In this chapter we attempt to address the issues discussed above by designing experiments in which participants have to generate explanations in a diagnostic task that is more complex than in previously reported studies and that at the same time are controlled enough to study memory effects. During this diagnostic reasoning task, we assess the availability of explanations not only at the end of a trial, but we also track the availability while new symptoms are observed. We do this with an implicit probe reaction time measure, rather than with an explicit measure of the explanations' availability. This should reduce potential effects of the measurement on the outcome of the task itself. Before we present the experiments in detail, we discuss the potential role of memory activation for the generation and evaluation of explanations.

Memory Activation and the Generation and Evaluation of Explanations

To understand the role of memory activation in diagnostic reasoning, it is necessary to consider how diagnostic knowledge is represented in memory (Arocha et al., 2005). A large number of studies have shown that with increasing experience in a domain reasoners develop knowledge structures whose content reflects the structure of the environment (Anderson & Schooler, 1991; Gigerenzer, Hoffrage, & Kleinbölting, 1991). To illustrate this using our earlier example, a physician will have a stronger memory representation of a diagnosis that has occurred frequently in the past, compared with a rare diagnosis. Similarly, the association between symptoms and their potential diagnoses in memory will increase with increasing experience of their co-occurrence. Given such a highly adapted knowledge structure, data extracted from the environment can serve as a cue for the retrieval of diagnostic hypotheses from long-term memory (Arocha & Patel, 1995; Ericsson & Kintsch, 1995; Kintsch, 1998; Thomas et al., 2008). An observation's efficiency as retrieval cue will depend on how strongly it is linked to the explanation in memory; the stronger the link, the more activation will occur (Anderson et al., 1998).

So far, we have looked at the question of how observed information can serve as a retrieval cue for one associated explanation from memory. However, a key aspect of diagnostic reasoning is that pieces of information are usually associated with a large number of possible explanations. Retrieving them all from memory is often impossible due to constraints set by cognitive capacity and the time available for diagnosis. To understand diagnostic reasoning it is therefore necessary to understand not only how one potential explanation is retrieved from memory but also how a selection is made among all the possible alternatives. For selecting explanations from a set of alternatives it is necessary to evaluate the alternatives in the set. A factor commonly linked to the evaluation of explanations is their coherence with the data. In his *theory of explanatory*

coherence, Thagard (1989a, 1989b, 2000) showed how a set of potential explanations can be evaluated purely on the basis of the coherence between the explanations and the observed data. In the computational implementation of this theory, ECHO, pieces of information are represented by interconnected nodes that, depending on their coherence to each other, spread activation or inhibition. The theory predicts that explanations most coherent with the observed data are most strongly available (because they receive a large amount of activation) and that explanations that are associated with only some of the observations have a lower availability (because they receive some inhibition). Applied successfully to explain phenomena in various domains, the theory has been described as a "computationally efficient approximation to probabilistic reasoning" (Thagard, 2000, p. 95). However, in its original implementation it is used to model the integration of information only at a certain point of time.

An extension of Thagard's theory that can account for sequential information integration has been proposed by Wang at al. (2006b; see also Mehlhorn & Jahn, 2009). They assumed that activation and inhibition spreading from new observations would add to the activation of observations that were observed before. Referring to work on memory retention, they proposed that the impact of observations decays exponentially with the square root of time. Consequently, over time observations should increasingly lose their impact on memory activation. This assumption is in contrast to recent findings that suggest that information in working memory seems to be subject to very little decay (Berman, Jonides, & Lewis, 2009; Jonides et al., 2008; Oberauer & Lewandowsky, 2008) or even no decay (Lewandowsky, Oberauer, & Brown, 2009). Thus, whereas constraint satisfaction seems to be a plausible mechanism for information integration at a certain point in time, the integration over time leaves open questions. Furthermore, the implementation of the theory into a connectionist network makes it difficult to assess how such a hypothesis evaluation mechanism would interact with the constraints set by other aspects of cognition, like perception, memory, and deliberate decision strategies.

A theory that takes into account the effect of limited cognitive resources on hypothesis generation and evaluation has recently been proposed by Thomas et al. (2008). In their HyGene model, diagnostic reasoning is described as a two-stage process, where a phase of automatic memory retrieval of hypotheses is followed by a phase of deliberate hypothesis evaluation. The memory retrieval stage itself consists of two parts. The first stage is a prototype extraction process, in which a memory trace is derived from episodic memory that "resembles those hypotheses that are most commonly (and strongly) associated with the data" (Dougherty et al., 2010, p. 308). In the second stage, this prototype is matched against known hypotheses in semantic memory. If sufficiently activated by the prototype, hypotheses from semantic memory are placed in working memory where they can be evaluated by deliberate reasoning processes. Although the authors stressed the importance of understanding sequential information integration and discussed possible related questions, they did not present predictions for the sequential integration of information. Such predictions are complicated due to the assumptions of two distinct memory systems that are involved in hypothesis generation. Would, for example, new observations lead to the retrieval

of different prototypes from memory? And if so, what would be the effects on the availability of hypotheses that were activated by previously retrieved prototypes?

Given the open questions presented above, we were interested in whether memory activation can indeed explain the generation and evaluation of explanations as found in an experimental setting. To answer this question, we extracted the most essential elements of the theories presented above and implemented them into a general cognitive architecture. For avoiding additional questions that might arise from understanding the interaction of episodic and semantic memory we focus on the effects on semantic memory. The basic assumption of the theories mentioned above is that each observation can affect the availability of explanations in memory. If an observation supports a particular explanation, the observation will spread activation to the explanation and will make it more available to the reasoner. If an observation does not support a particular explanation, the observation will spread inhibition to the explanation and make it less available to the reasoner.[1] If an observation is completely unrelated to an explanation, the explanation's activation will not be affected. Following the idea of Wang et al. (2006b), we assume that if several observations are currently in the focus of attention (that is, stored in working memory) they can serve as a sort of "combined retrieval cue" for explanations in long-term memory.

Current Chapter

As mentioned above, not much progress has been made in understanding how exactly sequentially made observations will affect memory activation over time. To shed light on this question, we implemented four different cognitive models. These models all share the general assumptions about memory activation and inhibition as presented above, but they vary with regard to how strongly newly incoming observations affect the availability of explanations over time. In a first model, *model-current*, at each point in time only the most recent observation affects the availability of explanations. This model is designed to test whether the assumption that sequentially observed symptoms serve as combined retrieval cue is necessary, or whether the activation and inhibition spread by the current symptom alone can fit the activation curves found in the experiments. In the remaining three models the observations serve as combined retrieval cue and, thus, all affect the explanations' availability. The models vary with regard to how strong each observation is weighed. One of the models, *model-time*, tests the assumption that observations are weighed according to the times since they were observed, as proposed by Wang et al. (2006b). As decay of information in working memory has been questioned (Berman et al., 2009; Jonides et al., 2008; Lewandowsky et al., 2009; Oberauer & Lewandowsky, 2008), we implemented two alternative models in which observed information does not decay. One model, *model-constant*, tests the assumption that observations are weighed according to the total amount of

[1] *In contrast to the concept of spreading activation between positively associated memory elements, the concept of spreading inhibition between negatively associated memory elements is neglected in many theories of memory retrieval, as it often has little practical impact (cf. Anderson & Lebiere, 1998). However, in diagnostic reasoning making a certain observation does not only increase the probability for positively associated explanations being the correct diagnosis, but it also decreases the probability of other explanations. Consequently, inhibition between observations and nonsupported explanations becomes important (Dougherty & Sprenger, 2006).*

information that is currently held in working memory. This assumption arises from the idea that the total amount of activation that can be spread from working memory is a limited and constant amount that will be equally divided between the elements in working memory (Lovett, Daily, & Reder, 2000). The fourth model, *model-number*, tests the assumption that all observations currently stored in working memory are weighed equally, independent of the time since they were observed and independent of the number of observations. Consequently, in this model, the total amount of activation and inhibition spread into long-term memory will increase with the number of observed symptoms.

To test the models, we conducted two behavioral experiments. In the experiments, participants had to find diagnoses for sequentially presented series of medical symptoms. The knowledge necessary to solve this task consisted of a number of symptoms, each of which was associated with a number of alternative explanations. Whereas the symptoms were real medical conditions, their association with the explanations was artificial to avoid possible effects of prior knowledge and for being able to fully balance the material. To be able to investigate the effects of memory activation on the generation of explanations, we tried to minimize the role of deliberate hypothesis evaluation strategies in the task. Therefore, experimental trials were generated in a way such that in most trials to find the correct diagnosis it was sufficient to retrieve the one explanation from memory that was most coherent to the set of observed symptoms. Thus, although each of the serially presented symptoms had a number of possible explanations that should vary in their availability over the course of the trial, at the end of the trial the most active explanation would also be the correct diagnosis. We expected the activation of explanations in memory to depend upon the serially observed symptoms as described above, with supporting symptoms increasing an explanation's availability and nonsupporting symptoms decreasing its availability. Activation was measured with a probe reaction task. The idea behind this task is based on lexical decision tasks where participants respond faster to a probe that is more highly activated in memory than to a probe of lower activation (e.g., Meyer & Schvaneveldt, 1971). We now first describe the method and the data from Experiment 1. Then we describe the cognitive models in detail and present the model results. Subsequently, we present Experiment 2, compare its results to predictions of the models, and discuss the implications of our findings.

Experiment 1

The goal of Experiment 1 was to test whether the availability of explanations over the course of diagnostic reasoning indeed depends upon the information observed over time. Therefore, we tracked the activation of three different kinds of memory elements during trials of a diagnostic reasoning task: (a) explanations that were supported by all the observed symptoms (compatible explanations), (b) explanations that were not supported by all of the observed symptoms (incompatible explanations), and (c) explanations that were completely unrelated to the symptoms (foils). (See the

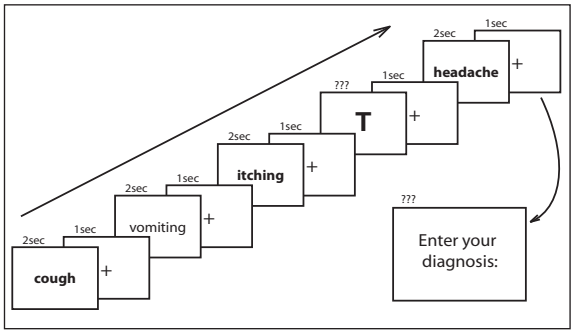


Figure 2.1 Illustration of the trial procedure for a sample trial from Experiment 1.

experimental-material section below for a more detailed description of the different kinds of explanations.) If the availability of explanations in memory depends on the observed symptoms as described above, we would expect symptoms to increase the activation of compatible explanations and to decrease the activation of incompatible explanations. The availability of foils should not be affected by the observed symptoms.

To introduce some uncertainty in the task, we varied the reliability of the symptoms presented in each trial. Whereas in 75% of the trials each of the symptoms reliably pointed towards the correct diagnosis (coherent trials), in 25% of the trials a misleading symptom was added that did not correspond to the correct diagnosis (incoherent trials). Participants were not told whether a trial was coherent or incoherent.

Method

Participants

Twenty-three undergraduate students from the Chemnitz University of Technology took part in this experiment. Of those, one participant had to be excluded from analysis, because she did not reach the required performance in the training session. Twelve of the remaining 22 students were female. The mean age was 24.1 (*SD* = 6.8).

Tasks

Diagnosis task. Participants were told that the main task they had to solve was to diagnose hypothetical patients after a "chemical accident". In each experimental trial, a set of three to four symptoms was presented and the chemical that explained the combination of these symptoms had to be found (see Figure 2.1 for a sample trial). This task allowed us to assess overall performance in the trials.

Probe task. The second task to be solved in the experiment was a probe task. After one of the symptoms in each trial, a probe was presented. Participants had to decide

Table 2.1 Domain knowledge participants had to acquire before Experiment 1.

Aggregate state and source of contamination	Category	Chemical	Specific symptoms		Unspecific symptoms		
Gasiform, inhaled	Landin	B	Cough	Shortness of breath	Headache		
		T	Cough	Vomiting	Headache		Itching
		W	Cough			Eye inflammation	Itching
Crystalline, skin contact	Amid	Q	Skin irritation	Redness	Headache		
		M	Skin irritation	Shortness of breath	Headache		Itching
		G	Skin irritation			Eye inflammation	Itching
Liquid, drinking water	Fenton	K	Diarrhea	Vomiting	Headache		
		H	Diarrhea	Redness	Headache		Itching
		P	Diarrhea			Eye inflammation	Itching

Note. Original materials were presented in German.

as fast as possible whether the probe (e.g., T in Figure 2.1) was the name of one of the chemicals learned in the training session (see Table 2.1) or not. Participants were told that the two tasks were not related to each other. This task allowed us to track the availability of explanations over the course of the diagnosis task.

Material

Learning material. The material that participants had to learn before the experiment consisted of nine different chemicals (see Table 2.1). Chemicals were named with single letters, which allowed us to construct balanced, artificial connections between symptoms and explanations about which participants would have no prior knowledge. Furthermore, using single letters as chemical names allowed us to use letters in the probe task, avoiding potential problems associated with the use of whole words (e.g., individual differences in reading speed and word frequency effects). The chemicals were grouped into the three artificial categories Landin, Amid, and Fenton. Participants were told that chemicals from the three categories differed in their state of aggregation: Landin chemicals, for example, were gasiform and affected especially the respiratory system because they were inhaled. This organization of knowledge into a hierarchical structure was used to ease the learning of the material by allowing participants to connect it to their knowledge about the biological workings of the human body. It reflects in a simplified form the hierarchical knowledge organization found in medical diagnosis (Arocha & Patel, 1995). Each chemical caused three to four medical symptoms. Symptoms had either a relatively small number of two or three

Table 2.2 Coherent and incoherent sample trials for Experiment 1.

Order	Symptoms	Explanations supported by current symptom	Possible target probes		Possible foils
			Compatible	Incompatible	
Coherent trial – Correct diagnosis: T					
1st	Cough	BTW	BTW	QMGKHP	FZVDNCXLR
2nd	Vomiting	TK	T	QMGKHP	FZVDNCXLR
3rd	Itching	TWMGHP	T	QMGKHP	FZVDNCXLR
4th	Headache	BTQMKH	T	QMGKHP	FZVDNCXLR
Incoherent trial – Correct diagnosis: B					
1st	Cough	BTW	BTW	QMGKHP	FZVDNCXLR
2nd	Eye inflammation	WGP	W	QMGKHP	FZVDNCXLR
3rd	Shortness of breath	BM	B	QMGKHP	FZVDNCXLR
4th	Headache	BTQMKH	B	QMGKHP	FZVDNCXLR

Note. Shown for each symptom are supported explanations, possible target probes, and foils. Note that the set of potential incompatible probes stayed the same over the trial (it consisted of those explanations that were not supported by the first symptom), whereas the set of potential compatible probes changed as the number of explanations supported by all symptoms decreased.

explanations (specific symptoms like cough) or a larger number of six explanations (unspecific symptoms like headache). This variance in the number of explanations was introduced because it is an important feature of real-world diagnostic knowledge that increases the complexity of the task.

Experimental material. Coherent trials were generated by presenting the three or four symptoms caused by one of the chemicals. In those trials all symptoms pointed coherently toward the correct diagnosis. Incoherent trials were generated by inserting an additional misleading symptom into the symptoms of one of the three-symptom chemicals (see Table 2.2 for a coherent and an incoherent sample trial). Apart from this manipulation, the order in which symptoms were presented in each trial and the order of trials were randomly chosen for each participant. Each diagnosis occurred with equal frequency during the experiment. Participants were told that, throughout the experiment, the second symptom of each trial might be misleading.[2] To keep them aware of this, the second symptom of each trial was printed in normal letters, whereas all other symptoms were printed in bold letters. Participants had no means of distinguishing coherent from incoherent trials until they observed the third symptom

[2] *Although manipulating uncertainty in this way represents a strong simplification of real-life diagnostic uncertainty, we chose this design for two main reasons. First, varying the position of the unreliable information within trials would have required a far larger number of trials. The number of trials already being very large, we decided against this (potentially very interesting) manipulation. Second, not informing participants about the potential unreliability of the second symptom might have resulted in a variety of potential strategies in dealing with incoherent trials (see Chinn & Brewer, 1998 for an overview of potential strategies in dealing with incoherent data). By informing participants which symptom might be unreliable, we attempted to reduce the amount of possible strategies.*

of the trial, which was either coherent with the second symptom (coherent trials) or not (incoherent trials).

To track the activation of explanations, a probe was presented after one of the symptoms in each trial. Each probe was a single letter that was either a target probe (one of the names of the nine chemicals) or a foil (see Table 2.2 for examples of the different probe types). Target probes were either *compatible targets* or *incompatible targets*.[3] Compatible targets probed explanations that were supported by all the symptoms preceding the probe (except for the misleading symptom in incoherent trials). Incompatible targets probed explanations that were not supported by all symptoms. The incompatible targets were chosen such that they were not supported by at least the first symptom of the trial. This allowed us to test the possible effect of inhibition beginning directly after the first symptom, where explanations that were supported by the symptom (compatible targets) could be compared to explanations that were not supported (incompatible targets). *Foils* were randomly sampled from nine letters that were not associated with any of the symptoms (see Table 2.2).

The type of probe (compatible target, incompatible target, or foil) and the position of the probe in the trial (after the first, second, third, or fourth symptom) were randomized over trials, with the constraints that (a) target probes and foils appeared equally often and (b) probes of each type appeared with equal frequency at all the positions. In 8.3% of the trials no probe was presented. Instead, after one of the symptoms of those trials, participants were asked to provide the set of diagnoses they currently had in mind. These 'no-probe' trials were intended merely to prevent participants from expecting a probe in each trial and were not analyzed.

Procedure

Each participant completed 5 sessions, which took part over a maximum of 10 days, with the first and second session on consecutive days.

Training session. The first session was a training session to ensure a high familiarity with the material and the task. It consisted of several blocks that were repeated until participants solved them with at least 80% accuracy. First, participants were presented with the cover story "diagnose patient after chemical accident" and with the complete knowledge (see Table 2.1). After a paper-and-pencil exercise in which they could use the table to write down which chemicals were associated with each symptom, participants had to study each chemical category separately on the screen. They were asked to memorize and report the name of the category, of the chemicals, and their respective symptoms. When they could report complete knowledge of the category at least once without error, they completed two more training blocks for that category. In the first block, sets of symptoms were displayed on the screen, and participants had to enter the chemical that caused this set of symptoms. In the second block, symptoms

[3] *In incoherent trials a third type of target probe was used (rejected targets). Rejected targets probed explanations that were compatible with early symptoms but incompatible with later symptoms. The reactions to those probes were in line with our predictions. However, as those probes were presented only in the incoherent trials, we will not report them here.*

were presented sequentially on the screen. After each symptom, participants were asked to enter all chemicals from the currently practiced category that could explain the symptoms seen so far.

After the training blocks for the single categories were completed, participants could again study the complete material (see Table 2.1). They were then presented with four training blocks for the complete material. The first block was identical to the final one in the single category training, but now all categories were tested. The second block was used to familiarize participants with the concept of incoherent trials; that is, they learned that the second symptom of each trial might be misleading. In the third block the probe task was introduced. After an explanation of the task, participants were presented with probes and had to decide whether they were targets (chemicals) or foils. The last block consisted of trials identical to the trials in the experiment. Participants were sequentially presented with symptoms. After one of the symptoms they had to react to a probe, and after all symptoms had been presented, they were asked for their diagnosis. Depending on a participant's performance, this session lasted between 60 and 90 min.

Experimental sessions. The experimental phase was split into four sessions. Each session began with a short practice block to refresh the participants' knowledge of the material. Afterwards participants solved 96 diagnostic reasoning trials, of which 75% were coherent and 25% were incoherent. The completion of the experimental trials in each session took about 30 min. Each trial was started self-paced. The symptoms of the trial were presented sequentially in the middle of the screen for 2 s each, with a fixation cross presented for 1 s in between (see Figure 2.1). After one of the symptoms in each trial, either the probe or the question for the current set of explanations was presented. The probe appeared in the form of a letter, and participants had to indicate if the letter was the name of a chemical by pressing a button on a response box. At the end of each trial, participants were asked to enter their diagnosis on a standard keyboard. Participants were instructed to solve the diagnosis and the probe task as accurately and fast as possible. Reactions times for probes and diagnoses were recorded from the moment that the probe/question for diagnosis appeared on the screen. After each input participants received feedback about their response accuracy.

Results

Probe reactions

To test the activation of explanations during the diagnostic reasoning trials, reaction times of correct probe responses were analyzed in coherent and incoherent trials with correct diagnoses. Scores above and below 3 *SD* from the condition mean of each participant were excluded from analysis, resulting in the elimination of 1.7% of the correct probe responses. [4]

[4] *To test for the robustness of our findings, we also conducted all analyses of the reaction time data based on the medians (without excluding outlier values). The primary results are consistent across analyses.*

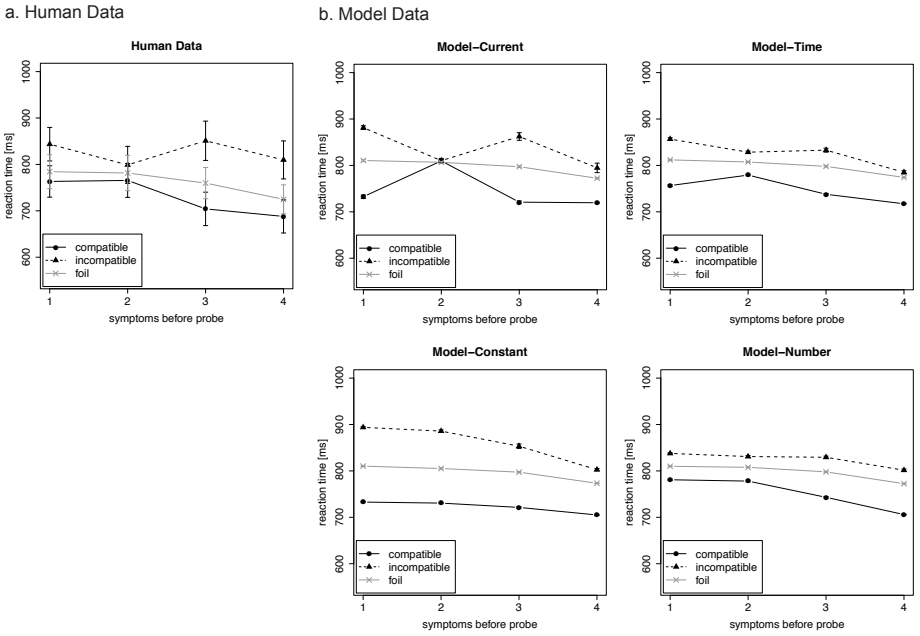


Figure 2.2 Mean (±1 *SE*) reaction time to probes over the course of trials in Experiment 1, showing (a) human data and (b) model data. The models are described later in the text.

Coherent versus incoherent trials. To test if the reaction time patterns differed depending on whether the third and fourth symptoms were coherent (coherent trials) or incoherent (incoherent trials) with the second symptom, we conducted an ANOVA[5] with the factors coherence (coherent vs. incoherent trial) and type of probe (compatible target, incompatible target, or foil). Symptoms before probe (three vs. four) was used as a numerical regressor variable. Neither the main effect of coherence, $F(1,21) = 1.642$, $p = .214$, $\eta_p^2 = .073$, nor any of the interactions involving coherence were significant: coherence × type of probe: $F(2,42) = 2.776$, $p = .074$, $\eta_p^2 = .117$; coherence × symptoms before probe: $F(1,21) < 1$; coherence × type of probe × symptoms before probe: $F(2,42) < 1$. Consequently, for further analyses we collapsed the data over the factor coherence.

Compatible versus incompatible versus foil. Figure 2.2a shows the reaction times of the different probe types over the course of the trials. Table 2.3 shows the results of the ANOVAs performed to analyze this data. First, an ANOVA with the factor type of probe (compatible target, incompatible target, or foil) and the numerical regressor symptoms before probe (one, two, three, or four) confirmed a significant interaction. To check whether this interaction was indeed caused by different slopes of all probe types, we conducted additional ANOVAs for each pair of probe types. They confirmed significant interactions for each pair, except for the pair compatible-foil. For this pair,

[5] *All ANOVAs were repeated-measures ANOVAs.*

Table 2.3 Results of the ANOVAs for compatible targets, incompatible targets, and foils after each symptom in Experiment 1.

Effect	Factors	F	p	η_p^2
Interaction	Type of probe (compatible, incompatible, foil) × Symptoms before probe (one, two, three, four)	(2,42) = 5.03	**.011**	.19
Interaction	Type of probe (compatible, incompatible) × Symptoms before probe (one, two, three, four)	(1,21) = 7.15	**.014**	.25
Interaction	Type of probe (compatible, foil) × Symptoms before probe (one, two, three, four)	(1,21) = 2.35	.140	.10
Main effect	Type of probe (compatible, foil)	(1,21) = 4.49	**.046**	.18
Interaction	Type of probe (incompatible, foil) × Symptoms before probe (one, two, three, four)	(1,21) = 3.70	**.068**	.15
Simple effect for compatible	Symptoms before probe (one, two, three, four)	(1,21) = 20.21	< **.001**	.49
Simple effect for incompatible	Symptoms before probe (one, two, three, four)	(1,21) = 0.46	.506	.02
Simple effect for foil	Symptoms before probe (one, two, three, four)	(1,21) = 25.56	< **.001**	.55
Simple effect after symptom 1	Type of probe (compatible, incompatible, foil)	(2,42) = 12.49	< **.001**	.37
Simple effect after symptom 2	Type of probe (compatible, incompatible, foil)	(2,42) = 1.21	.309	.05
Simple effect after symptom 3	Type of probe (compatible, incompatible, foil)	(2,42) = 29.13	< **.001**	.58
Simple effect after symptom 4	Type of probe (compatible, incompatible, foil)	(2,42) = 17.41	< **.001**	.45

Note. p values <.1 are shown in bold. For nonsignificant interactions the main effect of type of probe is also reported.

we additionally looked at the main effect of probe type, which showed to be significant, confirming that compatible probes are reacted to faster than foils. To test the course of availability over the course of the trial in more detail, we conducted additional simple effects analyses for each probe type. They confirm decreasing reaction times for compatible probes and foils. Incompatible probes did not vary over the course of the trial. Finally, simple effects analyses for symptoms before probe revealed significant differences between the probe types after all but the second symptom of the trial.

Diagnoses

To assess participants' performance in the diagnosis task, we measured diagnosis accuracy and diagnosis time at the end of each trial. For the analysis of diagnosis time, wrong diagnoses and diagnoses exceeding 3 *SDs* from the condition mean of each participant were excluded (resulting in an exclusion of 1.8% of the correct diagnoses). Diagnosis accuracy was equally high in coherent trials (M = 95.5%; SD = 4.1) and in incoherent trials (M = 95.5%; SD = 4.0), $t(21) < 1$. The equivalence between the

conditions was supported by a Bayes factor (BF) *t*-test, which showed clear evidence in favor of the null hypothesis ($BF = 6.13$).[6] This shows that the participants could solve the task well and, again, that there was no effect of a trial's coherence. Participants' time for entering correct diagnoses was fast overall but was significantly slower in coherent ($M = 795$ ms; $SD = 211$) than in incoherent trials ($M = 496$ ms; $SD = 125$), $t(21) = 8.612$, $p < .001$.[7]

Discussion

The results of the probe reaction task in Experiment 1 support the assumption that the availability of explanations over the course of diagnostic reasoning depends on the observed symptoms. Compatible targets (explanations supported by all symptoms) were responded to faster than incompatible targets (explanations not supported by all symptoms) and foils (not related to any symptom). This is in line with the prediction that explanations in memory receive activation from symptoms that support them. Incompatible targets were responded to not only slower than compatible targets but also slower than foils. This is in line with the prediction that symptoms inhibit explanations that they do not support.

An unexpected result of the probe reaction task was that the reaction times not only to compatible targets decreased over the course of the trial but also those to foils. Foils were letters that did not name chemicals and were therefore not related to any of the symptoms. Given a pure memory activation account, these letters should not change in their level of activation over the course of the trial, as they receive no activation or inhibition from any of the observed symptoms. A possible reason for the unexpected reaction time decrease might lie in our methodology. By presenting the probes with equal frequency after one of the four symptoms, we might have caused participants to be increasingly prepared to respond to the probe toward the end of the trial. Such an increasing response preparedness can be described by a hazard function (Chechile, 2003) and is comparable to the foreperiod effect (Vallesi, Shallice, & Walsh, 2007). The foreperiod effect is "usually observed when a range of variable FPs [foreperiods] occur randomly and equiprobably, [and] consists of reaction times (RTs) decreasing as the FP increases" (Vallesi et al., 2007, p.466). In our experiments, participants knew that after one of the symptoms in almost every trial a probe would appear. The position of the probes' occurrence was randomly and equiprobably distributed over the trials. With each symptom that went without a following probe, the likelihood for a probe increased. Participants could thus prepare for the probe and react slightly faster to it later on in the trial. Consequently, it is likely that part of the increase in response times

[6] *Bayes factors larger than 1.0 are taken as evidence in factor of the null, whereas Bayes factors less than 1.0 are taken as evidence in favor of the alternative. See Rouder, Speckman, Sun, Morey, and Iverson (2009) for derivations and a guide for interpreting the magnitude of Bayes Factors.*

[7] *Although the result of higher diagnosis times in coherent trials might seem counterintuitive, it is most likely caused by the number of symptoms presented before the diagnosis, rather than by the coherence of the trial. Incoherent trials always consisted of four symptoms, whereas coherent trials could consist of three (56% of all coherent trials) or four (44% of all coherent trials) symptoms. Analyzing coherent three-symptom and four-symptom trials separately shows that coherent four-symptom trials were in general responded to faster than coherent three-symptom trials ($M_{four} = 644$ ms, $SD = 208$; $M_{three} = 915$ ms, $SD = 223$, $t(21) = 11.233$, $p < .001$, and that the diagnosis times in coherent four-symptom trials were significantly faster than in incoherent trials, $t(21) = 4.684$, $p < .001$.*

to all probe types is caused by an increasing response preparedness over the course of the trial.

The manipulation of the symptoms' coherence affected neither the probe reaction times nor the accuracy of diagnoses. As explained above, participants could determine the correct diagnosis in incoherent trials by remembering that the second symptom of each trial is potentially misleading. A very simple strategy to use this knowledge would be to simply ignore the second symptom of each trial. Whereas such a strategy would lead to good performance in the incoherent trials and in most coherent trials, it would lead to suboptimal performance in a small part of the coherent trials, where ignoring the second symptom does not allow for unambiguously identifying the correct diagnosis (this was the case in 15% of the coherent trials). Nevertheless, a closer look at the probe reaction data seems to support such a strategy. Whereas reaction times differ significantly between the different probe types after the first, third, and fourth symptoms, they do not differ after the second symptom.[8]

Although the probe reaction time patterns are in line with our predictions, the comparison between verbal hypotheses and empirical data is usually reduced to a qualitative *descriptive level*. To test if memory activation, combined with ignoring the misleading symptom and increasing response preparedness over the trial, can also quantitatively *explain* the data, we developed computational cognitive models of the task. The models entail (a) the assumptions about memory retrieval as described in the introduction, as well as (b) the strategy to ignore potentially misleading information and (c) the participants' increasing preparedness to respond over the trial.[9]

Models

Model Description

To reach maximum comparability between the models, we implemented them all within one modeling framework, the cognitive architecture ACT-R (Anderson et al., 2004). From all the variants of potential modeling accounts we chose ACT-R because it puts a strong emphasis on processes underlying memory activation (Anderson et al., 1998; Anderson & Schooler, 1991) and integrates these processes with general assumptions about human cognition. It accounts for both subsymbolic and symbolic components of cognition and, therefore, allows for the implementation

[8] *To further test if participants indeed ignored the second symptom, we compared the diagnostic performance in coherent trials where ignoring the second symptom allowed for unambiguously finding the correct diagnosis (unambiguous coherent trials) and coherent trials where ignoring the second symptom did not allow for finding the correct diagnosis (ambiguous coherent trials). Indeed diagnosis accuracy was marginally higher in unambiguous ($M = 95.8\%$; $SD = 3.8$) than in ambiguous coherent trials ($M = 93.8\%$; $SD = 7.4$), $t(21)=1.815$, $p = .084$. Diagnosis times for correct diagnoses were considerably faster in unambiguous ($M = 757$ ms; $SD = 195$) than in ambiguous coherent trials ($M = 1053$ ms; $SD = 363$), $t(21)=5.297$, $p < .001$, suggesting that participants used time at the end of the trial to solve the ambiguity caused by ignoring the second symptom.*

[9] *Building the ignoring of misleading information and the increasing response preparedness into the models allowed us to assess whether the response pattern indeed could have been caused by the interaction of memory activation and these task- specific factors. It is important to note however, that these additional model components alone would not have been able to fit the participants' responses. Without an effect of observations on reaction times to the probes, ignoring the second symptom would not predict any effect on the reaction time data by itself. Increasing response preparedness alone would predict a decrease of reaction times over the trial, but no differences or interactions between the different probe types.*

of automatic memory processes as well as deliberate reasoning strategies and their possible interaction. It has received empirical support and validation from a large number of studies in a variety of research areas (ranging from simple list memory tasks, Anderson et al., 1998; to language acquisition, Taatgen & Anderson, 2002; see http://act-r.psy.cmu.edu for an extended list of publications). Furthermore, ACT-R allows for modeling of the complete task, as solved by the participant. Thereby, without requiring additional assumptions about how the model maps on the experiment, it produces results that are directly comparable to human data. This is possible because the ACT-R theory predicts not only the probability and latency of retrieving facts from declarative memory but also the time taken to perceive a stimulus and give a response (e.g., by pressing a key).

Knowledge about facts is represented in the form of chunks in ACT-R's long-term memory, which is commonly referred to as declarative memory. Chunks can represent observations (e.g., medical symptoms), as well as their potential explanations (e.g., medical diagnoses). Access to the chunks depends on their activation in memory (Anderson, 2007; Lovett et al., 2000). Only chunks whose activation exceeds a certain amount, the *retrieval threshold,* τ, can be retrieved. The probability, p, that a chunk i will cross the retrieval threshold, τ, depends on its activation, A_i:

$$p = \frac{1}{1 + e^{\frac{\tau - A_i}{s}}} \tag{2.1}$$

where s reflects the amount of noise added to the chunk's activation.

If a chunk i is activated strongly enough to be retrieved, its activation, A_i, determines the time required for the retrieval. The more active the chunk, the faster it can be retrieved. The time it takes to retrieve chunk i is a negative exponential function of its activation, A_i, as shown in Equation 2.2:

$$Time = Fe^{-A_i} \tag{2.2}$$

where F is a parameter scaling the latency of retrievals.

The idea behind the concept of a chunk i's activation, A_i, is that the strength of activation reflects the likelihood (specifically, the log odds) of the chunk being needed in the near future (Anderson & Schooler, 1991). This likelihood is determined by three factors: the chunk's usefulness in the past, B_i, its usefulness in the current context, S_i, and a random noise component, ε:

$$A_i = B_i + S_i + \varepsilon \tag{2.3}$$

The chunk's usefulness in the past is reflected by the base-level activation, B_i. ACT-R predicts that the more often a chunk has been retrieved from memory and the more recent these retrievals were, the higher its activation. This prediction can explain empirical findings that show that explanations with high base-rates of occurrence are generated more often and earlier than explanations with low base-rates. Although the effects of an explanation's previous use are an interesting aspect of memory effects

in diagnostic reasoning, they are not the focus of the current chapter. Therefore, base levels were kept at a constant level in the model. This was plausible because participants received extensive training on the task (leading to a saturation effect) and all symptoms and explanations appeared equally often in the experiment.

The important factor for our research question is the second part of Equation 2.3: the chunk's usefulness in the current context, S_i. A chunk's usefulness in the current context reflects the likelihood that the chunk will be needed given the information currently available from the environment. In diagnostic reasoning, the current context is defined by the to-be-explained observations (e.g., the medical symptoms displayed by a patient; Arocha et al., 2005; T. R. Johnson & Krems, 2001; Thomas et al., 2008). ACT-R predicts that an explanation *i* that is stored in long-term memory receives activation, S_i, from each observation *j* that is currently stored in working memory:[10]

$$S_i = \sum_j W_j S_{ji} \tag{2.4}$$

where the amount of spreading activation, S_i, is determined by the associative strength, S_{ji}, between explanation *i* and observation *j,* scaled by the amount of activation that can be spread from working memory, W_j. As we describe in detail below, we manipulated this scaling parameter, W_j, to implement different ways of sequential information integration in the different models. The associative strength, S_{ji}, represents the extent to which observation *j* increases or reduces the likelihood that the explanation *i* is needed from memory. This relationship can be described by a log conditional probability ratio (Anderson & Lebiere, 1998):

$$S_{ji} = \log \frac{p(observation_j \mid explanation_i)}{p(observation_j \mid \text{not}(explanation_i))} \tag{2.5}$$

where the numerator describes the probability that observation *j* has been observed when explanation *i* is needed (i.e., is valid in this context) and the denominator describes the probability that *j* has been observed when *i* is not needed. Using an example, the equation describes the probability for observing the symptom cough while having the flu divided by the probability for observing cough while not having the flu. As the likelihood to observe cough is higher when having the flu than when not having the flu, Equation 2.5 predicts a positive associative strength between cough and flu. In contrast, if an observation (cough) does not support an explanation (pregnancy), the likelihood to observe cough when the patient is pregnant decreases. This results in a negative associative strength.

Although Equation 2.5 provides a good estimate for associative strengths between chunks, their exact calculation is often computationally intractable (Anderson & Lebiere, 1998). Following ACT-R, we approximate positive associative strengths, S_{ji}, between chunks as

$$S_{ji} = S - \ln(fan_{ji}) \tag{2.6}$$

[10] *To model working memory we use one of the buffers of ACT-R's cognitive modules, the imaginal buffer. The imaginal buffer is commonly used to hold a mental representation of the problem currently in the focus of attention (Borst et al., 2010).*

where S is a parameter for the maximum associative strength between chunks in memory and fan_{ji} is the number of chunks i that are positively associated with a chunk j. Following this equation, an observation that is associated with only few explanations (e.g., a medical symptom that is specific to a certain group of diseases) has a lower fan and therefore a higher associative strength to the explanations than an observation that is associated with many explanations (e.g., a medical symptom that is associated with a variety of diseases). Although the associative strength between positively associated symptom-explanation pairs can be estimated as shown in Equation 2.6, the estimation of "negative associations" is problematic. Depending on the certainty that is assumed in the task, the values for S_{ji} resulting from Equation 2.5 would lie somewhere between $-\infty$ (if it is absolutely certain that an explanation can be excluded from consideration when a certain observation is made) and 0 (if it is not known whether a certain observation and explanation can occur together). As ACT-R provides no solution for this issue, we treat negative associative strengths as a free parameter that we estimate from our empirical data.

Four Different Models of Sequential Information Integration

To implement the different assumptions of how observations might affect the availability of explanations over time, we used the parameter W_j. This parameter scales the amount of activation and inhibition that each observed symptom can spread to long-term memory. For reaching maximum comparability between the models, we kept the total amount of W after the fourth symptom at a constant level between the models.[11] Consequently, in all four models, the same amount of activation is spread from working memory after all symptoms have been observed. The models vary in how this activation is distributed amongst the symptoms and in how it varies over the course of the trial in the following ways.

Model-current

In the first model, at each point in the trial, only the most recently observed symptom spreads activation and inhibition to explanations in long-term memory. We implemented that by setting W_j for all but the current observation to zero. The current observation was scaled with value W.

[11] *For being able to directly compare the levels of explanations' availability over the course of the trial, we kept the total amount of the scaling parameter W constant after the fourth symptom of the trial. This choice was somewhat arbitrary, as we could have kept W constant at any other point during the trial (e.g., using a constant value W_j after the first symptom of the trials). Note however, that this would have not changed the results substantially, as it would have merely produced a linear transformation of all scaling values. To test this we implementing all models with a constant value of W_j=.16. This produced the same pattern over the course of the trial, however, with much smaller differences between the different probe types at each point during the trial, leading to much smaller values for R^2 and lower diagnosis accuracies for all models except model-number.*

Model-time

In the second model, all observed symptoms spread activation and inhibition. As proposed by Wang et al. (2006b), the amount of activation spread by each of the symptoms depends on the time since the observation was made. The most recently observed symptom is weighed most strongly. Earlier observations are weighed with a decayed strength, with the strength decaying exponentially in the square root of time:

$$W_j = W_{j-1}(1-d)^{\sqrt{t}} \tag{2.7}$$

Model-constant

In the third model, all observed symptoms spread activation and inhibition. As proposed by Lovett et al. (2000), the total amount of activation that can be spread from working memory has a constant value W. If several observations j are stored in working memory, they share this total activation. Consequently, the more symptoms are observed, the smaller is the impact of each of these symptoms:

$$W_j = \frac{W}{n} \tag{2.8}$$

Model-number

In the fourth model, the total amount of activation spread from working memory at a certain point in time depends on the number of observed symptoms. Each symptom can spread a fixed amount of activation, resulting in an increasing amount of spreading activation and inhibition with an increasing amount of observed symptoms. Consequently, in this model the amount of activation spread by each of the observations neither depends on the time since the observation was made nor on the number of observations. Each symptom is scaled with the same value W_j.

Model Procedure

All models follow the same procedure, with the only difference between the models being the setting of parameter W as described above. The model code can be downloaded from http://www.ai.rug.nl/~katja/models. As for the participants in our experiments, the models observe sequentially presented medical symptoms, diagnose the chemical that caused these symptoms, and react to the probe that is presented after one of the symptoms. The knowledge necessary to solve this task (see Table 2.1) is represented in the models' declarative memory and consists of two different types of facts, represented as chunks. The first type reflects the possible symptoms. The second type represents the letters that can be presented during the experiment (chemicals and foils) and their associated information. Each letter is represented by a chunk that holds

the letter's name, the information stating whether it is a chemical or a foil, and, for chemicals, the associated symptoms.[12]

When a symptom is presented on the screen, the model moves its attention to the symptom, reads it, and retrieves its meaning from declarative memory. The symptom is then stored in working memory. This process is repeated for each observed symptom so that, over the course of a trial, working memory is successively filled with the observed symptoms. Stored in working memory, symptoms automatically spread activation and inhibition to explanations in declarative memory as described by Equation 2.4. To simulate the strategy of ignoring the potentially misleading symptom, the second symptom observed in each trial is not stored in working memory. When the question for the final diagnosis is presented on the screen, the model retrieves that explanation from declarative memory that receives the most activation from the symptoms in working memory and enters the respective letter. The letter representing the correct explanation is most strongly associated with the observed symptoms. However, as described above, the different models vary in how the associative strength between the symptoms and their explanations are weighed. In *model-current*, only the current symptom spreads activation. Thus, at the point of diagnosis, only the last of the observed symptoms affects activation of explanations in memory. In the remaining models all observed symptoms spread activation at the point of diagnosis. In *model-time*, the strength of activation depends on the time since an observation was made. Consequently, even though all observations affect explanations' availability in memory, availability is most strongly affected by newer observations. In *model-constant* and *model-number*, at the point of diagnosis, each symptom is weighed with equal strength. As the letter representing the correct explanation is most coherent with the symptoms, it obtains the highest amount of spreading activation and is the one most likely to be retrieved. However, as shown in Equation 2.3, due to random noise also in these models it can happen that an alternative explanation receives more activation and is incorrectly entered as diagnosis.

When a probe is presented, the models move their attention to the probe and retrieve the chunk representing the probe letter. If that letter is stored as a chemical, the models respond "yes"; if it is stored as a foil, the models respond "no". As described by Equation 2.2, the speed with which a chunk can be retrieved depends on its activation. The more spreading activation the chunk receives from the symptoms in working memory, the higher it will be activated and the faster the retrieval. Thus, as in human participants, the time the models need to respond to a probe can be used as a measure of the activation of explanations in memory. To simulate the participants' increasing response preparedness over the trial, the models retrieve expectations about whether the upcoming stimulus is a symptom or a probe. If the retrieved expectation is met by the presented stimulus, the stimulus is processed as explained above. If the expectation is violated, the models need to make a change to their expectation before they can process the stimulus. This change in expectation costs 50 ms. The later in the

[12] *Note that not only the chemicals but also the foils are represented in memory. This is because, contrary to lexical decision tasks, where a constrained number of words stands against an unconstrained number of non-words, in our experiment chemicals and foils each consisted of a set of nine letters which were taught to the participants in the training session.*

Table 2.4 Fits for probe reaction times (R^2 and RMSD) and diagnostic performance (Diagnosis Accuracy and Diagnosis Time) of each model for Experiments 1 and 2.

	R^2	RMSD (ms)	Diagnosis Accuracy (%)	Diagnosis Time (ms)
Experiment 1				
Human Data, M (SD)			95.5 (3.7)	705 (167)
Model-current	.79	30	28	586
Model-time	.79	28	53	597
Model-constant	.70	38	86	592
Model-number	**.85**	**27**	**85**	**569**
Experiment 2				
Human Data, M (SD)			95.9 (3.9)	574 (264)
Model-current	.24	61	27	566
Model-time	.37	75	71	584
Model-constant	.45	60	95	587
Model-number	**.71**	**83**	**92**	**589**

Note. The best fitting model is indicated in bold. RMSD = root-mean-square-deviation.

trial the probe is presented, the higher the chance that it is expected by the models and that no time-costly expectation changes have to be made.[13]

Results and Discussion of the Models

The models were run for each participant on the trials that this participant had solved. As described above, the four different models varied in their setting of the values for the parameter, W_j, that weighs the strength of observations j in working memory. All other parameters were kept constant between the models. To fit the models, we estimated the speed and stochasticity of memory retrievals, the base-level activation of facts in memory, and the amount of spreading activation from symptoms to explanations.[14] All other parameters were kept at the default values of ACT-R 6.0 (Anderson, 2007).

Following the analysis of the human data, we collapsed the models' data over the factor coherence. The resulting reaction times to the probes are shown in Figure 2.2b.

[13] *Reflecting the probabilities for upcoming stimuli, the base-level activations of the expectations vary. As probes are presented equally often after one of the four symptoms, the probability of a probe being presented after the first symptom is only .25. Consequently, the base level of an expect-probe chunk after the first symptom is so much lower than the base level of an expect-symptom chunk that the model will retrieve an expect-probe chunk only in about 25% of all trials. With each additional symptom that is presented without a probe, the probability of a probe (reflected by the base levels of the expect-probe chunks) increases (to .33, .5, and 1 respectively). Consequently, the earlier in the trial the probe appears, the higher the chance that the model retrieves no expect-probe chunk and has to make a time-costly change to its expectation. The model changes its expectation by firing an additional production rule (costing 50 ms).*

[14] *ACT-R's latency factor (F) was set to 1.4 and activation noise (s) to .05. All facts in memory were set to equal, relatively high base levels of 2, modeling trained participants. Positive associative strengths (S_{ji}) were calculated using Equation 2.6, with the maximum associative strength (S) set to 2.5. Negative associative strengths (S_{ji}) were estimated from the data to be –.75. The total amount of W that the models spread after four symptoms were presented was set to .48.*

Fits for the probe reaction times and the diagnostic performance reached by each model are shown in Table 2.4. All models produce the basic result that, overall, compatible probes are reacted to fastest. This happens because in all models compatible probes receive more activation from the observed symptoms than do all other probe types. In all models, incompatible probes are slower than or at about the same level as foils. This happens, because in all models incompatible probes receive inhibition as well as activation from the observed symptoms. The reaction times to foils over the course of the trial are identical in all models, because these reaction times are not affected by spreading activation. As in the human data, they decrease over the trial. In the models this decrease is solely caused by the varying expectations about upcoming stimuli, suggesting that part of the decrease of reaction times to all probes was indeed caused by an increasing preparedness to respond. All models produce comparable diagnosis times. The models differ in the course of activation for compatible and incompatible probes and in the accuracy of their diagnoses in the following ways.

Model-current

Merely using the current symptom at each point in time, the model produces a surprisingly good fit to the probe reaction pattern. The model produces no difference between probe types after the second symptom, because no activation and inhibition is spread to long-term memory at this point. After all other symptoms, reaction times for compatible probes are faster than for foils because compatible probes receive activation from the current symptom. However, contrary to the human data, reaction times to compatible targets do not increase over the course of the trial. Incompatible probes are slower than foils, with a decrease of reaction times over the course of the trial. This happens because incompatible explanations are explanations that are incompatible to at least the first symptom of the trial. Consequently, incompatible probes always receive inhibition from the first symptoms, and they can receive inhibition, as well as activation, from the later symptoms. The model has poor diagnostic performance, which is not surprising, as in this model only the last symptom of the trial affects activation of explanations at the point of diagnosis.

Model-time

Letting the impact of observed symptoms decay over time, the model produces a good fit to the empirical probe reaction data. After the second symptom the difference between probe types is smallest, because at this point in the trial, only the decayed activation and inhibition of the first symptom affect explanations' availability. After all other symptoms, reaction times to compatible probes are faster and decrease over the course of the trial as the amount of spreading activation increases with each observed symptom. However, this decrease is much less pronounced than in the human data. Reaction times to incompatible probes also decrease, because the later in the trial, the higher the chance that incompatible probes not only receive inhibition but also activation from the observed symptoms. The model produces correct diagnoses in

about half of the trials, because in this model, symptoms that are presented late in the trial have an overproportional impact on explanations' availability.

Model-constant

By letting the observations at each point in time share a constant amount of total working-memory activation, this model also produces a good overall fit. However, here the visual inspection of the time course of explanations' activation also shows some deviations from the human data. In the model, at each point in time a constant amount of activation is spread from working memory. Consequently, compatible explanations stay at a constant level over the course of the trial (with a slight decrease caused by increasing response preparedness over the trial). Incompatible explanations stay at a constant and relatively high level of reaction times between the first and the second symptom and then decrease considerably. The model produces a high proportion of correct diagnosis, which is only slightly lower than in the empirical data.

Model-number

By increasing the amount of spreading activation and inhibition with each observed symptom, the model produces the best overall fit to the human data. As in the human data, reaction times to compatible probes do not change from the first to the second symptom and do decrease afterwards. This happens because compatible probes receive an increasing amount of activation with all but the second symptom. Incompatible probes slightly decrease over the course of the trial as they receive inhibition as well as activation. Like *model-constant*, the model does not reproduce the dip in reaction times to incompatible probes after the second symptom. The model produces the same proportion of correct diagnoses as *model-constant*, because after the last symptom of the trial they are identical due to the setting of the total amount of parameter W at this point.

To summarize, all models produce the overall pattern of probe response times as found in the human data. The models vary in how well they fit details of activation levels over the course of the trials. Only *model-constant* and *model-number* are able to produce high diagnostic performance, because they weigh all symptoms with equal strength at the point of diagnosis. However, even these models underpredict the diagnosis accuracy as well as the diagnosis times found in the human data. This underprediction is caused by the fact that in part of the coherent trials, ignoring the second symptom does not allow for finding a correct diagnosis. Whereas, as discussed earlier, participants might try to remember the second symptom once they realize that they cannot distinguish between explanations otherwise, the models do not have such knowledge. When simply relying on memory activation they have no means to correctly distinguish between alternatives if they receive an equal amount of activation from the observed symptoms. This result is a good illustration of the importance of automatic memory activation to interact with deliberate reasoning. Whereas in most experimental trials it was sufficient to enter the diagnosis suggested by memory

activation, in coherent trials where ignoring the second symptom led to equal activation of alternatives, participants most likely used additional deliberate reasoning processes to find the correct explanation.

In the experiment, participants had to diagnose coherent and incoherent sets of symptoms, because we wanted to add uncertainty to the task and because we were interested in seeing what happens in cases where memory activation alone might not be sufficient to find the correct explanation. As the empirical and model data for diagnoses and probe reactions suggest, participants dealt with that challenge by simply ignoring the potentially misleading symptom. They did so even though they were told to use all the presented symptoms for their diagnosis, they were trained to do so in the practice session, the information was misleading in only 25% of the trials, and ignoring the second symptom reduced diagnosis performance in 15% of the coherent trials. As suggested by the probe reaction data and the models, using this strategy was highly adaptive, because it allowed for finding the correct diagnosis by simply relying on memory activation in the vast majority of the trials.

Experiment 2

Experiment 2 had three main goals. First, we wanted to test the reliability of the key findings from Experiment 1 with an experimental setup that allowed us more control over participants' strategies. Therefore, symptoms in this experiment always coherently pointed toward the correct diagnosis. During trials we again tracked the activation of compatible explanations (supported by all symptoms), and incompatible explanations (not supported by at least the first symptom), and foils (not related to the symptoms). Second, we wanted to investigate in more detail the availability of explanations that are associated with only part of the symptoms observed in the trials. Therefore, in this experiment we tracked the availability of an additional group of explanations: rejected explanations. These explanations are supported by the initial symptoms of a trial but not by symptoms presented later on in the sequence. Consequently, they have to be rejected from the set of potential explanations at some point in the trial. Being able to inhibit such no-longer-compatible explanations has been described as one of the crucial aspects of diagnostic performance (Dougherty & Sprenger, 2006). To assess the activation of rejected explanations over the course of the task, we compared the activation of explanations that were (a) rejected at different points in the trial and (b) measured at different time spans after rejection. Third, we wanted to test how well the different models generalized to a new data set.

Method

Participants

Twenty-nine undergraduate students from the Chemnitz University of Technology who did not participate in Experiment 1 took part in this experiment. Three of them

Table 2.5 Domain knowledge participants had to acquire before Experiment 2.

Aggregate state and source of contamination	Category	Chemical	Specific symptoms		Unspecific symptoms		
Gasiform, inhaled	Landin	B	Cough	Shortness of breath	Headache	Eye inflammation	
		T	Cough	Shortness of breath	Headache		Itching
		W	Cough			Eye inflammation	Itching
Crystalline, skin contact	Amid	Q	Skin irritation	Redness	Headache	Eye inflammation	
		M	Skin irritation	Redness	Headache		Itching
		G	Skin irritation			Eye inflammation	Itching
Liquid, drinking water	Fenton	K	Diarrhea	Vomiting	Headache	Eye inflammation	
		H	Diarrhea	Vomiting	Headache		Itching
		P	Diarrhea			Eye inflammation	Itching

Note. Original materials were presented in German.

had to be excluded from data analysis, as they did not reach the required performance in the training phase. The resulting 16 female and 10 male participants had a mean age of 22.8 (*SD* = 3.6).

Material

Training material. The material that participants had to acquire in the training phase (see Table 2.5) was a slightly modified version of the material from Experiment 1. Again, chemicals were grouped into categories and caused three or four symptoms each. Whereas in Experiment 1 each symptom was caused by chemicals of either one, two, or all three categories, symptoms in this experiment were caused either by chemicals of only one category (specific symptoms like cough) or by chemicals of all three categories (unspecific symptoms like headache).

Experimental material. In the experimental phase participants solved trials that were comparable to the coherent trials of Experiment 1 (see Table 2.6 for a sample trial). The only difference was that now rejected explanations were also probed. These explanations varied in the point of their rejection during the trial and in the number of symptoms presented between the rejection and the respective probe. This manipulation resulted in three different types of rejected target probes: *rejected-after-2*, which could be presented after the second, third, or fourth symptom; *rejected-after-3*, which could

Table 2.6 Sample trial for Experiment 2.

Order	Symptoms	Explanations supported by current symptom	Possible target probes				
			Compatible	In-compatible	Rejected-after-2	Rejected-after-3	Rejected-after-4
Correct diagnosis: T							
1st	Headache	BTQMKH	BTQMKH	WGP			
2nd	Cough	BTW	BT	WGP	QMKH		
3rd	Shortness of breath	BT	BT	WGP	QMKH	-	
4th	Itching	TWMGHP	T	WGP	QMKH	-	B

Note. Shown for each symptom are supported explanations and possible target probes. Dashes indicated where cells cannot be filled in this particular trial. Foils were identical to those in Experiment 1.

be presented after the third or fourth symptom; and *rejected-after-4*, which could only be presented after the fourth symptom. This allowed us to investigate not only the course of an explanation's activation after its rejection but also the potential effect of the point when it is rejected in the trial. To prevent participants from expecting a probe in each trial, in 14% of the trials no probe was presented, but instead the question for the current diagnosis was asked after one of the symptoms. Again, these 'no-probe' trials were not analyzed.

Procedure

The experiment consisted of one training session and two experimental sessions. In both experimental sessions participants solved 170 diagnostic reasoning trials, with a 5-min break after half of the trials were completed. Except for this, the procedure was identical to Experiment 1.

Models

To generate predictions for the data of this experiment we used the models as described above, with the only change being that the models now did not ignore the second symptom of the trial. Except for the total amount of memory activation that was increased to reflect the higher number of observed symptoms in the trial, none of the parameters of the model were changed.[15]

[15] *As no symptoms were ignored, the models now had one more symptom to integrate than in Experiment 1. To account for this, we adjusted the setting of parameter W. The total amount of W that the models spread after four symptoms were presented was set to .64.*

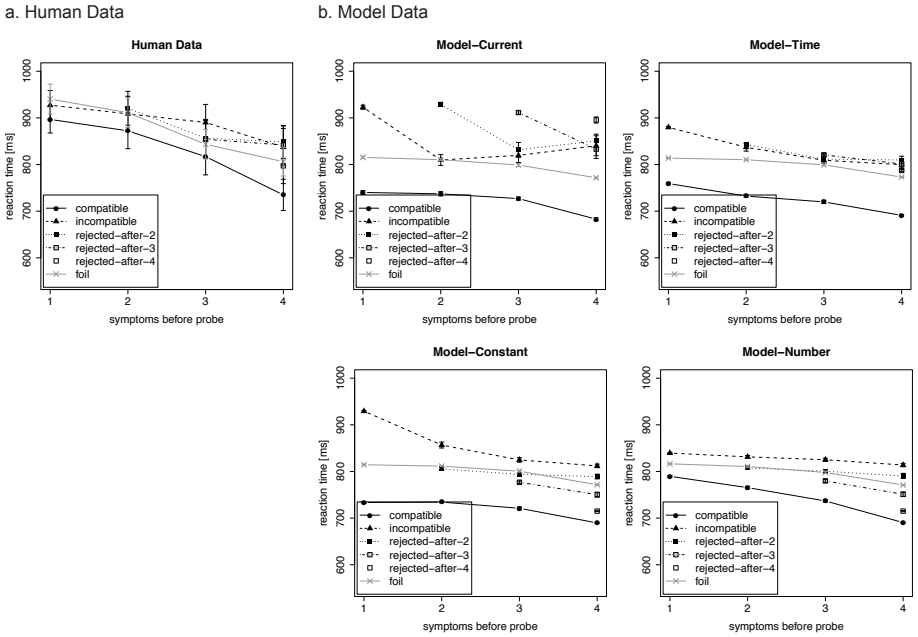


Figure 2.3 Mean (±1 *SE*) reaction time to probes over the course of trials in Experiment 2, showing (a) human data and (b) model data.

Results

Probe reactions

Reaction times of correct probe responses were analyzed in trials with correct final diagnoses. Scores above and below 3 *SDs* from the condition mean of each participant were excluded from data analysis, resulting in the elimination of 2.0% of the correct probe responses. The reaction times to all types of probes are presented in Figure 2.3a. Due to the incomplete design, analyzing the data with standard analyses is difficult. Here we present analyses for three subsets of the data that are most interesting to test our predictions. Subsequently we present the model fits, which cover the complete data set.

Compatible versus incompatible versus foil

First, we tested whether our results for compatible and incompatible target probes and foils could be replicated. Therefore, we did the same analyses as in Experiment 1; detailed results of the corresponding ANOVAs are shown in Table 2.7. An ANOVA with the factor type of probe (compatible target, incompatible target, or foil) and the numerical regressor symptoms before probe (one, two, three, or four) confirmed a significant interaction. To check whether this interaction was indeed caused by different slopes of all probe types, we conducted additional ANOVAs for each pair of

Table 2.7 Results of the ANOVAs for compatible targets, incompatible targets, and foils after each symptom in Experiment 2.

Effect	Factors	F	p	η_p^2
Interaction	Type of probe (compatible, incompatible, foil) × Symptoms before probe (one, two, three, four)	(2,50) = 3.84	**.028**	.13
Interaction	Type of probe (compatible, incompatible) × Symptoms before probe (one, two, three, four)	(1,25) = 5.65	**.025**	.19
Interaction	Type of probe (compatible, foil) × Symptoms before probe (one, two, three, four)	(1,25) = 0.90	.352	.04
Main effect	Type of probe (compatible, foil)	(1,25) = 10.88	**.003**	.30
Interaction	Type of probe (incompatible, foil) × Symptoms before probe (one, two, three, four)	(1,25) = 3.39	**.077**	.12
Simple effect for compatible	Symptoms before probe (one, two, three, four)	(1,25) = 34.46	< **.001**	.58
Simple effect for incompatible	Symptoms before probe (one, two, three, four)	(1,25) = 9.49	**.005**	.28
Simple effect for foil	Symptoms before probe (one, two, three, four)	(1,25) = 68.46	< **.001**	.73
Simple effect after symptom 1	Type of probe (compatible, incompatible, foil)	(2,50) = 4.37	**.018**	.15
Simple effect after symptom 2	Type of probe (compatible, incompatible, foil)	(2,50) = 2.10	.133	.08
Simple effect after symptom 3	Type of probe (compatible, incompatible, foil)	(2,50) = 3.76	**.030**	.13
Simple effect after symptom 4	Type of probe (compatible, incompatible, foil)	(2,50) = 7.60	**.001**	.23

Note. p values <.1 are shown in bold. For nonsignificant interactions the main effect of type of probe is also reported.

probe types. As in Experiment 1, they confirmed significant interactions for each pair, except for the pair compatible-foil. For this pair, we additionally looked at the main effect, which again was significant, confirming that compatible probes are reacted to faster than foils. To test the course of availability over the course of the trial in more detail, we conducted additional simple effects analyses for each probe type. They showed that, for all probe types, reaction times decrease over the course of the trial. Finally, simple effects analyses for the symptoms before probe revealed significant differences between the probe types after all but the second symptom of the trial.

Compatible versus incompatible versus rejected-after-2 versus foil

To test how the activation of rejected explanations changes with time after their rejection, we analyzed the course of activation of explanations that were rejected after the second symptom. Detailed results of the corresponding ANOVAs are shown in Table 2.8. An ANOVA with the factor type of probe (compatible, incompatible, rejected-after-2, and foil) and the numerical regressor symptoms before probe (two,

Table 2.8 Results of the ANOVAs for rejected-after-2 targets, compatible targets, incompatible targets, and foils after symptoms two, three, and, four in Experiment 2.

Effect	Factors	F	p	η_p^2
Interaction	Type of probe (rejected-after-2, compatible, incompatible, foil) × Symptoms before probe (two, three, four)	(3,75) = 1.89	.138	.07
Main effect	Type of probe (rejected-after-2, compatible, incompatible, foil)	(3,75) = 8.44	< **.001**	.25
Interaction	Type of probe (rejected-after-2, compatible) × Symptoms before probe (two, three, four)	(1,25) = 4.52	**.043**	.15
Interaction	Type of probe (rejected-after-2, Incompatible) × Symptoms before probe (two, three, four)	(1,25) < .01	.980	< .01
Main effect	Type of probe (rejected-after-2, Incompatible)	(1,25) = .06	.811	< .01
Interaction	Type of probe (rejected-after-2, foil) × Symptoms before probe (two, three, four)	(1,25) = 1.20	.284	.05
Main effect	Type of probe (rejected-after-2, foil)	(1,25) = .06	.149	.08
Simple effect for rejected-after-2	Symptoms before probe (two, three, four)	(1,25) = 5.80	**.024**	.19
Simple effect after symptom 2	Type of probe (rejected-after-2, compatible, incompatible, foil)	(3,75) = 2.09	.108	.08
Simple effect after symptom 3	Type of probe (rejected-after-2, compatible, incompatible, foil)	(3,75) = 2.70	**.052**	.10
Simple effect after symptom 4	Type of probe (rejected-after-2, compatible, incompatible, foil)	(3,75) = 6.91	< **.001**	.22

Note. p values <.1 are shown in bold. For nonsignificant interactions the main effect of type of probe is also reported.

three, or four) showed no overall interaction but a significant main effect of type of probe. To compare rejected-after-2 targets to each of the other probe types, we conducted additional pairwise ANOVAs. They reveal that rejected-after-2 targets interact with compatible targets, but do not interact with or differ from incompatible targets and foils. To test the course of availability of targets rejected-after-2 symptoms over the course of the trial, we conducted a simple effects ANOVA. It showed that reaction times for these targets also decrease over the course of the trial. Finally, simple effects analyses for reactions after two, three, and four symptoms revealed significant differences between the probe types after the third and fourth symptoms.

Time since rejection

The analysis of rejected-after-2 targets that is reported above sheds some light on the course of explanations' activation after rejection. However, a potential problem with this analysis is that it confounds the time since rejection and the time of measurement. Systematic effects of the time of measurement (e.g., the foreperiod effect or the number of compatible explanations at the point of testing) might thereby drown out the effects of the time since an explanation's rejection. Therefore, we conducted an additional

analysis in which we compared the different types of rejected targets (rejected-after-2, rejected-after-3, and rejected-after-4) when tested after the fourth symptom. An ANOVA with the factor type of probe (compatible, incompatible, foil, rejected-after-2, rejected-after-3, and rejected-after-4) confirmed that after the fourth symptom, reaction times differed significantly between the probe types, $F(5, 125) = 5.085$, $p < .001$, $\eta_p^2 = .169$. Holm-corrected pairwise comparisons showed that reactions to compatible targets were faster than reactions to all other probes ($p < .04$), except for probes rejected after the fourth symptom ($p = .172$). No other difference reached significance. This confirms the prediction that explanations supported by all symptoms receive the most activation and suggests that the activation of rejected targets indeed differs depending on the time since rejection.

Diagnoses

Again, we assessed accuracy and time for entering the diagnoses at the end of each trial. For the analysis of diagnosis times, wrong diagnoses and diagnoses above and below 3 *SDs* from the condition mean of each participant were excluded (resulting in an exclusion of 2.5% of correct diagnoses). The high diagnosis accuracy (95.9%; *SD* = 3.9) and short time for entering correct diagnoses (574 ms; *SD* = 264) show that participants could solve the diagnosis task with high performance.

Model predictions

Model predictions for the probe reaction times are presented in Figure 2.3b. The associated fits and the diagnostic performance reached by each model are shown in Table 2.4. The model that also produced the best fit in Experiment 1, *model-number,* generalizes best to the probe reaction data of Experiment 2. A visual inspection of the model predictions shows that this model predicts the time course of compatible and incompatible probes very well and better than the other three models do. For rejected probes the picture is less clear. *Model-constant* and *model-number* make almost identical predictions for rejected probes. Whereas these predictions are very good for rejected-after-4 probes, *model-time* seems to predict the time course of rejected-after-2 and rejected-after-3 probes better. However, in interpreting these results, it should be kept in mind that all predictions of the best fitting model, *model-number,* are within the standard errors of the empirical data. Again, only *model-constant* and *model-number* are able to produce the high diagnostic accuracy as found in the empirical data.

Discussion

Experiment 2 had three main goals: (a) to replicate the findings about the availability of compatible and incompatible explanations and foils in a more controlled setup, (b) to allow a closer evaluation of the availability of rejected explanations, and (c) to test how well the models generalize to a new data set. We were able to replicate the results for compatible and incompatible explanations. The inspection of rejected probes suggests

some difference between these probes, depending on the time since their rejection. The model comparison reveals large differences in generalizability of the models. *Model-number* predicts the probe reaction data time and the diagnostic performance well, whereas the remaining models show clear deviations from the data. *Model-number* is able to predict the effects for compatible and incompatible targets and foils. More interestingly, it is also able to approximate the pattern of the different types of rejected targets. The explanations rejected at different points in time had not been probed in Experiment 1 and therefore it was not self-evident that any of the models would be able to predict them.

Given that the parameters of the models were fit to Experiment 1 and not adjusted to the data of this experiment, the best fitting model, *model-number*, also reaches a lower fit in Experiment 2 than in Experiment 1. This is not surprising, as reaction times in Experiment 2 decreased more strongly than reaction times in Experiment 1. Reasons might be found not only in differences between the samples but also in differences between the tasks of the two experiments. In Experiment 1, participants had to keep in mind that symptoms might potentially be misleading and therefore that the current explanation might have to be changed during the trial. In Experiment 2, no such uncertainty existed and therefore participants could allocate more resources to the probe task. By adjusting parameters characterizing the sample (e.g., duration of memory retrievals) and the task (e.g., how strong response preparedness increases over the trial), the model could be fit to produce reaction times closer to those of the humans. In the current chapter we decided to forgo this adjustment, because we were interested in seeing how well the models generalize to a new data set (see Böhm & Mehlhorn, 2009, for earlier versions of the models that were fit to part of this data set). The fact that without parameter adjustment *model-number* was able to predict the major effects found in the human data lends additional support to this model, as the ability of a model to generalize to a new data set, without any further parameter adjustments, has been described as an important standard by which models should be evaluated (Marewski & Olsson, 2009; Pitt, Myung, & Zhang, 2002; Roberts & Pashler, 2000).

General Discussion

In diagnostic reasoning, reasoners have to generate and evaluate possible explanations for data observed from the environment. Whereas the number of potential explanations is often large, reasoners usually generate and deliberately evaluate only a small subset of explanations. Empirical research has shown that the selection of explanations into the generated subset seems to be highly adaptive to previous experience and the current reasoning context (Dougherty et al., 1997; Dougherty & Hunter, 2003a; Gettys et al., 1987; Sprenger & Dougherty, 2006; Weber et al., 1993). However, although the idea that currently available observations affect the generation of explanations from memory seems obvious, few studies have experimentally tested this assumption. Even less work has investigated how newly incoming observations affect the availability

of explanations over time. The goal of this chapter was to more closely investigate how automatic memory processes can provide the reasoner with an adaptive selection from memory over time. We report the results of two behavioral experiments that were designed to overcome potential problems of earlier studies. The results of the experiments are compared with predictions of four cognitive models. Implemented in the cognitive architecture ACT-R, these models test hypotheses about how sequentially observed information might affect the availability of explanations in memory over time.

In both experiments participants diagnosed quickly and with high accuracy. Whereas all models diagnosed equally fast, only the models that weighed each observation equally strongly at the point of diagnosis (*model-constant* and *model-number*) were able to replicate the high diagnosis accuracy. The models reached this performance by merely relying on spreading activation between symptoms and explanations, suggesting that, given sufficient knowledge, memory activation can indeed provide the reasoner with a highly adaptive selection of explanations from memory. The models' underprediction of diagnostic performance in trials of Experiment 1 where memory activation alone was not sufficient to find the correct diagnosis shows where deliberate reasoning processes might come into play.

The probe reaction task proved to be a useful measure for the availability of different explanations over the course of the reasoning task. Whereas for the participants the probe task seemed unrelated to the diagnosis task, reaction times to probes of different explanations varied, as predicted, as a function of the observed symptoms over time. All models were able to reproduce the overall activation differences between explanations found in the human data. This lends support to the basic assumption of spreading activation and inhibition as it was implemented in all models. The models differed in their ability to reproduce the courses of explanations' activation over time. In Experiment 1, all models reached a high overall fit, with varying success in fitting details of the activation curves. Furthermore, all models but *model-constant* reflected the ignoring of the second symptom in their curves. The generalization test of Experiment 2 shows that *model-number* generalizes best to the new data set. The success of this model suggests that the impact of observations on memory activation might depend neither on the time since an observation was made nor on the number of observations. Rather, the results suggest that all observations that are stored in working memory seem to be weighed equally at each point in time until an explanation is found.

Generalizing to Real-World Diagnostic Reasoning

To allow for the experimental control that was necessary to test our assumptions about memory activation, the experiments and models in this chapter present a simplified version of diagnostic reasoning. In real-world diagnostic reasoning, the task characteristics, the memory representation, and the reasoning strategies will often be more complex. This increased complexity raises a number of issues, which we briefly discuss here.

An important issue for understanding real-world diagnostic reasoning is the interaction of automatic processes as investigated here with more deliberate reasoning strategies. Our models assume a very simple strategy: Observed symptoms are successively stored in working memory and, when asked for the diagnosis, the explanation that receives the most activation from the observed symptoms is retrieved from memory. Obviously, such a simple strategy oversimplifies diagnostic reasoning. Whereas we chose to implement such a simple strategy to test different assumptions about automatic memory activation processes over time, it is very likely that people use additional deliberate strategies. People probably start to retrieve possible explanations early on in the reasoning process (see, e.g., Just & Carpenter, 1987, for evidence that people interpret evidence as soon it becomes available). Thus, presumably, not only are the sequentially acquired observations stored in working memory but also potential explanations that have been retrieved from long-term memory. Such an additional strategy of retrieving explanations earlier in the reasoning process might explain some of the deviation between our probe data and the model predictions. For example, all models underpredicted the decrease of the slope of reaction times for compatible targets in both experiments. If the reasoner additionally would retrieve candidate explanations and store them in working memory, these explanations would be available at low time cost. Therefore, mean reaction times to compatible targets would decrease over the course of the trial to a stronger extent than predicted by our pure activation-based models.

The question about reasoning strategies is closely linked to another important question for understanding real-world diagnostic reasoning. How do people represent the sequentially observed data and the generated explanations in working memory? As discussed above, for the sake of simplicity, in our models only observations are stored in memory. Storing observations is not implausible, as it has been found that not yet explained observations are kept in a more active state in memory than explained observations (Baumann, 2001). However, a more comprehensive account of diagnostic reasoning will also have to incorporate predictions about the representation of already retrieved explanations and their influence on memory activation over time.

A key aspect of such considerations has to be the contrast between limited human working-memory capacity and the large number of observations and explanations that might have to be maintained during diagnostic reasoning tasks. In our experiments participants had to maintain up to four symptoms in working memory, a number that lies within the accepted range of 4±1 (Cowan, 2001). However, assuming that participants also store retrieved candidate explanations in memory, one would quickly reach capacity limits. Furthermore, in most real-life diagnostic reasoning tasks, a higher number of observations needs to be explained. An interesting question for further research will be to investigate what happens if the amount of information to be actively maintained during the task exceeds working-memory capacity. In such a case, the least activated information might be dropped from working memory (Chuderski, Stettner, & Orzechowski, 2006; Thomas et al., 2008) and therefore should lose its ability to spread activation to long-term memory, unless it is actively recovered from long-term memory.

Also time and task constraints will be more complex in many real-world settings. In our experiments, symptoms were presented at a fixed rate, with a relatively small spacing over time, and with (almost) no interference from other tasks. It has been proposed that information will be held by a cognitive resource like working memory until the resource is needed for another task (Salvucci & Taatgen, 2008). Applied to diagnostic reasoning as proposed in this chapter, this would mean that observed symptoms would remain in working memory until working memory is needed for something else (see also Berman et al., 2009). With increasing spacing of the symptoms over time, and with increasing complexity of the diagnostic situation, the chance for interfering working-memory use grows. Consequently, the probability for observed symptoms to be lost from working memory also grows under these conditions. Also in this case, symptoms would have to be actively recovered from long-term memory before they could affect memory activation again.

Another open question is related to the representation of knowledge in long-term memory. As we discussed in the introduction, memory activation processes can only provide the reasoner with an adaptive set of possible explanations if diagnostic knowledge is represented in a way that fits the requirements of the task. Memory activation might for example favor the retrieval of an explanation that has been successfully used in the past compared with the retrieval of an explanation that has rarely occurred in the reasoner's experience but fits the current patient better. The representation of knowledge in long-term memory will most probably vary depending on the task structure and the way in which it was learned. In our experiments, the task structure was clearly defined, and the knowledge was learned in an explicit semantic fashion through a series of practice trials. This simplification of knowledge acquisition compared with real-life situations allowed us to focus on the effects of memory activation by keeping the effects of knowledge representation relatively constant. It will be an interesting question for future research to investigate the role of different ways of knowledge representation in memory activation processes. By proposing an episodic as well as a semantic representation and specifying the memory activation processes related to these representations, Thomas et al. (2008) already made an important step in this direction. We suspect, however, that a more detailed investigation of different ways of knowledge representation will not call the implications of our findings into question. A less clearly defined task structure and a more implicit acquisition of knowledge as they would be expected to occur in real-life will only increase the importance of memory activation processes (Dijksterhuis & Nordgren, 2006).

Conclusion

To conclude, our results support the assumption that automatic memory activation can adaptively regulate the availability of explanations in memory and thereby provide the reasoner with a subset of explanations that have a high probability of being relevant in the current context. This regulation of explanations' availability was evident not only at the point of the diagnosis but throughout the whole reasoning process. Future research must show whether simple models of memory activation as we tested them in this

chapter prove to be sufficient to explain memory processes in real-world diagnostic reasoning tasks. Further research is also needed to investigate how such simple memory models can be extended into more comprehensive models of diagnostic reasoning that take into account the interaction and respective contributions of automatic memory activation and deliberate reasoning strategies.

Modeling Information Integration with Parallel Constraint Satisfaction

In which I explore different methods of modeling sequential information integration with parallel constraint satisfaction.

An earlier version of this chapter was published as:

Mehlhorn, K. & Jahn, G. (2009). Modeling sequential information integration with parallel constraint satisfaction. In N.A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Modeling Information Integration with Parallel Constraint Satisfaction

Chapter

Abstract

An important aspect of human cognition is the sequential integration of observations while striving for a coherent mental representation. Recent research consistently stresses the importance of fast automatic processes for integrating information available at a certain point in time. However, it is not clear how such processes allow for maintaining a coherent and up to date mental representation in the light of new information. We compare variants of two methods of modeling sequential information integration with parallel constraint satisfaction models: (1) carrying over results from the previous integration step and (2) decaying input strength of older observations. Results of these models for coherent and incoherent sets of observations are compared to human data from a diagnostic reasoning task.

Introduction

A key feature of many everyday reasoning tasks is that observations are processed sequentially. Whether it is in diagnostic reasoning, in decision-making, or in belief updating, often information becomes available step by step. If a large amount of information is given all at once, it might only be perceived and understood sequentially due to limited cognitive capacities. Although possible implications of the sequential nature of tasks (e.g., order effects) have been discussed (e.g., Hogarth & Einhorn, 1992; Wang et al., 2006b), the underlying cognitive mechanisms are not fully understood. Recent research consistently points out the importance of fast automatic processes for integrating information available at a certain point in time (e.g., Glöckner & Betsch, 2008). However, it is not clear how such processes allow for maintaining a coherent mental representation in the light of new incoming information. In this chapter, we explore alternative implementations of such processes in connectionist parallel constraint satisfaction models.

Previous research has shown that reasoners hold knowledge structures that reflect the structure of the task in the environment (Anderson & Schooler, 1991; Gigerenzer et al., 1991). For example, a physician learns, with an increasing number of patients encountered, which symptoms are associated with which diseases and how strong these associations are. Given such an adapted knowledge structure, observations can serve as a cue for the retrieval of associated knowledge from long-term memory (e.g., Baumann et al., 2007; Kintsch, 1998; Thomas et al., 2008). To maintain a coherent representation of the task at hand, this newly activated information somehow needs to be integrated with previous observations and previously activated knowledge. How is this achieved?

Wang et al. (2006b) have proposed a connectionist model of sequential information integration based on the idea of explanatory coherence that, probably most prominently, was introduced by Thagard (1989a, 1989b, 2000) in the field of scientific discovery. Thagard implemented explanatory coherence among interconnected propositions in a connectionist constraint satisfaction model (ECHO). In ECHO, propositions are represented by nodes. The nodes are interconnected by symmetric excitatory and inhibitory links representing the relations (constraints) between them. Nodes representing observed information are additionally connected to a special activation node (special evidence unit = SEU), which always has an activation value of 1 and is the model's "energy source". Connecting not all, but only these data nodes to the energy source reflects the idea that empirical data are weighted more strongly than theoretical hypotheses held by the reasoner (Thagard, 1989a).

The strength of a proposition in the network is indicated by the numerical activation of its node. Before the network is integrated, activation of all nodes is set to default values. Then, activation spreads from the SEU to the data nodes and then to other connected nodes. The net input each node receives is calculated as the weighted sum of the activation of all nodes it is connected to. After calculating the input for each node, the activation of all nodes is updated synchronously. These two steps are repeated iteratively, until activation stops changing substantially. The more coherent

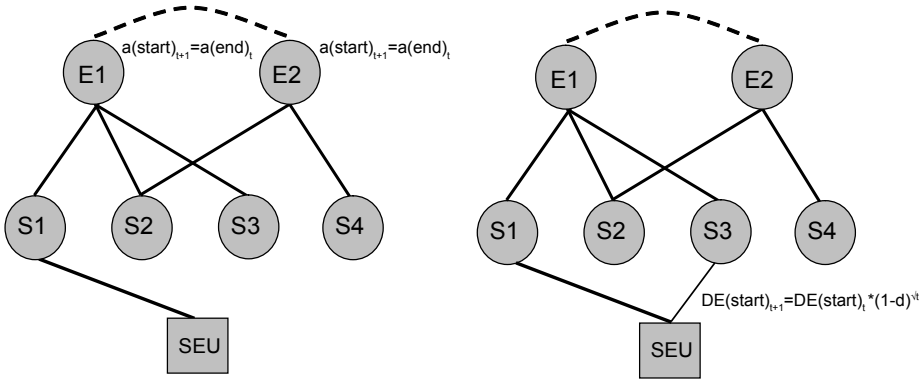


Figure 3.1 Two basic approaches to model sequential data in a constraint-satisfaction network. Either the previous state of the model is preserved by retaining the initial activation of the explanation nodes (left) or previous symptoms keep influencing the activation in the network by a (decaying) connection to the SEU (right).

a proposition is with the observed information and other related propositions, the higher is the activation of its node when the network settles.

The idea of constraint satisfaction has been widely applied to areas such as text comprehension (Kintsch, 1998), social impression formation (Thagard, Kunda, Read, & Miller, 1998), visuo-spatial reasoning (Thagard & Shelley, 1997), causal reasoning (Hagmeyer & Waldmann, 2002), medical diagnosis (Arocha & Patel, 1995), and decision making (Glöckner & Betsch, 2008). In all of these different tasks, reasoners need to find an interpretation that is more coherent with the available information than possible alternative interpretations. Such coherent interpretations can be the meaning of a word that fits best in the current context, the impression about a person that is most coherent with one's previous impression about him/her, or it can be the diagnosis that best explains the set of a patient's symptoms.

Applied successfully to model various phenomena in all the above domains, constraint satisfaction models have been described as a "computationally efficient approximation to probabilistic reasoning" (Thagard, 2000, p. 95). However, Thagard's ECHO has some major limitations. For our question most importantly, it only models the parallel integration of information given at a certain point in time. To incorporate newly incoming observations in a sequential task, a new network would have to be constructed.

Wang et al.'s UECHO (uncertainty-aware ECHO; 2006b), shares the basic features of ECHO, but can handle sequentially incoming observations. This is achieved by two basic changes. First, the network contains not only the currently available information as in ECHO, but all possible observations are included from the beginning. Thus, when new observations come in, the network does not have to be restructured. Second, the models differ with regard to which observations are connected to the special evidence unit (SEU). While in ECHO, all observation-nodes are connected to the SEU, in UECHO, only those nodes representing information observed up to the

current point in time are connected to the SEU. Due to these two changes, when a new piece of information is observed, the model does not have to be rebuilt, but only a new connection between that observation and the SEU needs to be added.

For modeling sequential information integration, it is not only important to incorporate new observations into the network, but also to coherently integrate this new information with the previous state of the network. One could think of two basic approaches for implementing this preservation of the previous state (illustrated in the networks in Figure 3.1). In both networks, the upper nodes, E1 and E2, represent possible explanations of the possible observed symptoms S1-S4 (represented by the nodes in the middle row). Solid lines between the nodes represent coherent relations (e.g., E1 explains S1), dashed lines represent incoherent relations (e.g., E1 and E2 contradict each other). In both networks, the symptoms S3 and S1 have been observed.

In the left network the previous state of the network is preserved by retaining the activation of the explanation nodes. When the first symptom (S3) is observed, S3 is connected to the SEU and the activation for the explanation nodes (E1 and E2) is calculated. The resulting activation values are used as starting values for the integration of the new symptom (S1).

The right network illustrates the approach proposed by Wang et al. (2006b). Here, the activation of all nodes is reset to default before each new run. The preservation of the previous state is obtained indirectly, by connecting not only the new information, but also previously observed information to the SEU. In the example, S3 as well as S1 are connected to the SEU. Therewith, the older observation (S3) can continue influencing the current activation in the network. To account for sequential observations, the strength of this influence decays over time. The most recently observed symptom (S1) gets a strong connection to the SEU, whereas older observations (S3) are connected to the SEU with a decayed strength. This strength (data excitation, DE) is a function of a decay rate *d* and the time interval since the symptom was observed. By referring to work on memory retention, Wang et al. (2006b) propose to let DE decay exponentially in the square root of time.

We will show that the first modeling alternative - retaining output activation from previous runs - is not appropriate for modeling the integration of sequential information, because of the dynamics of spreading activation in the network. The second alternative is explored in more detail. The resulting activation for both approaches is tested against human data.

Experiments

Design and Procedure

Human data on memory activation during sequential symptom integration was obtained in two diagnostic reasoning experiments: Experiment 1 (Mehlhorn et al., 2008) and Experiment 2 (Baumann et al., 2007). (For a more detailed description of the experiments, see also Chapter 2 of this thesis.) In these experiments, participants diagnosed hypothetical patients after a chemical accident. For each patient, a set of

Table 3.1 Domain Knowledge Participants had to Acquire Before Experiment 1.

Aggregate state and source of contamination	Category	Chemical	Specific symptoms		Unspecific symptoms		
Gasiform, inhaled	Landin	B	Cough	Shortness of breath	Headache		
		T	Cough	Vomiting	Headache		Itching
		W	Cough			Eye inflammation	Itching
Crystalline, skin contact	Amid	Q	Skin irritation	Redness	Headache		
		M	Skin irritation	Shortness of breath	Headache		Itching
		G	Skin irritation			Eye inflammation	Itching
Liquid, drinking water	Fenton	K	Diarrhea	Vomiting	Headache		
		H	Diarrhea	Redness	Headache		Itching
		P	Diarrhea			Eye inflammation	Itching

Note. Original materials were presented in German.

symptoms was presented sequentially on a computer screen and the task was to find the chemical that best explained the set of symptoms. The knowledge necessary to solve this task was taught to participants in an extensive training session. In both experiments, the knowledge consisted of nine different chemicals (named with single letters), grouped into three categories. Each chemical caused three to four symptoms. Symptoms were ambiguous, because each symptom could be caused by two to six different chemicals. Consequently, only the combination of symptoms allowed for unambiguously identifying the correct diagnosis (see Table 3.1 for the knowledge used in Experiment 1).

Two types of trials were used: In both experiments, coherent trials were presented. Additionally, in Experiment 1, incoherent trials were presented (see Figure 3.2 for a coherent and an incoherent sample trial). In coherent trials, all symptoms coherently pointed toward one explanation. Thus, the participants' initial explanation was supported by all later symptoms. In incoherent trials, the explanation suggested by the first two symptoms was incoherent with the later symptoms. Here, participants needed to revise their initial explanation after observing the third symptom. In such incoherent trials, it should be particularly difficult to integrate symptoms while maintaining a coherent mental representation. In total, in Experiment 1, participants were presented with 384 trials, of which 75% were coherent and 25% were incoherent. In Experiment 2, 340 trials were presented, which were all coherent.

In both experiments, two types of dependent measures were obtained. First, after all symptoms of a patient were presented, participants explicitly provided their diagnosis. Second, a probe reaction task was used as an implicit measure of the activation of

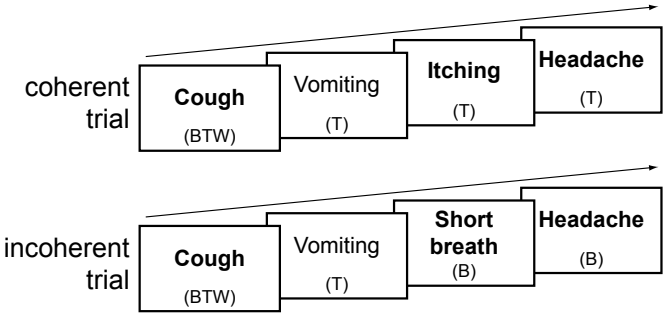


Figure 3.2 Example for a coherent and an incoherent trial in Experiment 1. Letters in parentheses show the compatible explanations after each symptom; they where not visible to the participants.

explanations during the sequential task. This measure is based on the idea of lexical decision tasks (Meyer & Schvaneveldt, 1971) according to which participants should respond faster to a probe that is highly activated in memory than to a probe of low activation. Each probe was a single letter that was either one of the names of the nine chemicals (targets) or one of nine other letters (foils). Participants were to decide as fast as possible whether the probe was a chemical name or not. To reduce possible influences of the probes on each other, only one probe was presented in each trial. Using this measure, it was possible to monitor the activation of explanations over the course of the sequential reasoning task with as little impact on the task itself as possible.

Such an implicit measure that directly tracks the activation of explanations in memory is especially suited to evaluate the validity of constraint satisfaction models. The usual approach to test these models is to compare the activation calculated in the model to an explicit measure obtained in human experiments. For example, Wang et al. (2006b) asked their participants for explicit belief ratings after each new observation. However, explicitly asking participants during the course of the task might influence the outcome of the task itself (Hogarth & Einhorn, 1992). Directly assessing the activation in memory with an implicit task reduces such a possible influence.

In this chapter, we use response times to target-probes (chemical names) for three different types of explanations to test the constraint satisfaction models. First, we are interested in explanations that are compatible with all symptoms observed before the probe's presentation (*compatible explanations*). Second, we are interested in explanations that are compatible with the initial symptoms, but that are incompatible with later symptoms (*rejected explanations*). Third, we look at explanations that are incompatible with at least the first symptom of the trial (*incompatible explanations*). Reactions to probes for the three kinds of explanations are compared at three different times of measurement over the course of the reasoning task (after *two*, *three*, and *four* symptoms). In Experiment 1 rejected explanations were only presented in incoherent trials. Therefore, below we report the data from the incoherent trials of Experiment 1 and compare them to the coherent trials from Experiment 2.

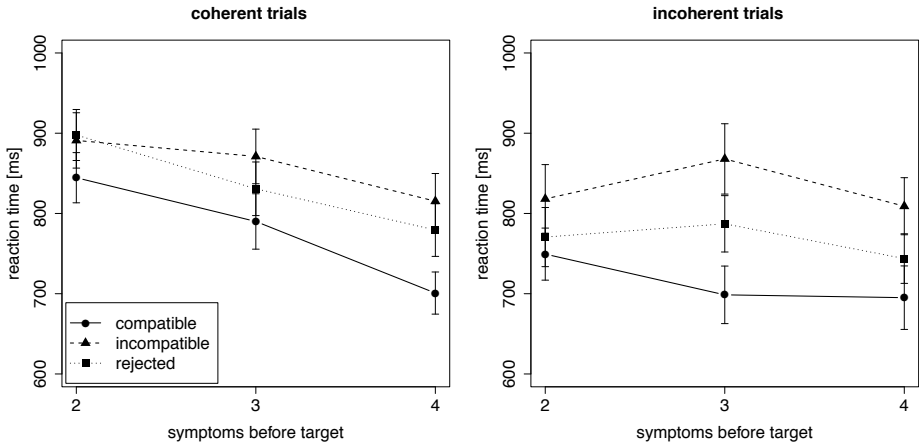


Figure 3.3 Reaction time to compatible, rejected, and incompatible target probes after the 2nd, 3rd, and 4th symptom. Left: coherent trials from Experiment 2 (Baumann et al., 2007); Right: incoherent trials from Experiment 1 (Mehlhorn et al., 2008).

Results

Diagnosis. In both experiments and in both types of trials, the accuracy of diagnoses given at the end of each trial was high (around 95%). This suggests that also in incoherent trials participants were able to solve the task easily.

Probe reaction task. In both types of trials, the fastest probe responses occurred for compatible explanations. Rejected explanations were responded to slower than compatible explanations, but faster than incompatible explanations (see Figure 3.3; see also Chapter 2 of this thesis for a more elaborate analysis of the behavioural data from the experiments). Coherent and incoherent trials differed in the course of activation over time. In coherent trials, reaction times decreased with an increasing number of symptoms, with the highest decrease for compatible explanations. In incoherent trials, this decrease was less visible, possibly because integrating the information was more difficult than in coherent trials. Nevertheless, the fast responses to compatible explanations suggest that, also in incoherent trials, participants managed to integrate the symptoms correctly.

Models

To assess the validity of the alternative modeling approaches, we implemented the knowledge used in the experiments into different constraint-satisfaction networks (see Figure 3.4 for an example). All networks consisted of the complete material participants needed to learn before the experiment. We used 9 nodes representing

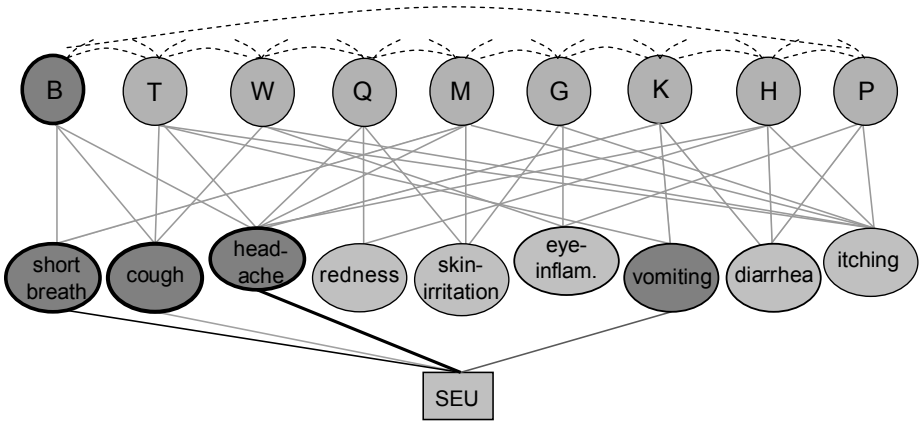


Figure 3.4 Network for an incoherent trial (cough - vomiting - short breath - headache) in Simulation 3. Dashed lines: inhibitory connections, solid lines: excitatory connections. B has the strongest activation when the network settles.

the symptoms, 9 nodes representing the explanations (chemicals), and connections representing the relations between those nodes. Nodes representing explanations were interconnected by inhibitory links, because the symptoms of each trial were caused by only one chemical. Symptoms were connected to their associated explanations by excitatory links.

In the networks, four basic parameters can be varied:

1. The initial activation of the explanation nodes before each run.
2. The initial activation of the symptom nodes before each run.
3. The strength of the connection between the nodes.
4. The strength of the connection between the symptom nodes and the special evidence unit (SEU).

To model the two basic approaches described above, we used variations of the parameters 1 and 4. The values of parameters 2 and 3 were set to fixed values: The initial activation of symptom nodes (parameter 2) was set to 1 for the currently observed symptom and to 0 for all other symptoms. The connection-strength between the nodes in the network (parameter 3) was set to 0.04 for excitatory and to -0.04 for inhibitory links.

To evaluate the models' capacity to emulate human information integration during the course of the task, we will now take a closer look at the process measure. For each model, we calculated the activation for the three types of explanations (compatible, rejected, and incompatible) at the three different times of measurement (after two, three, or four symptoms). This activation is compared to the human probe reaction-time data, which indicates memory activation of explanations.

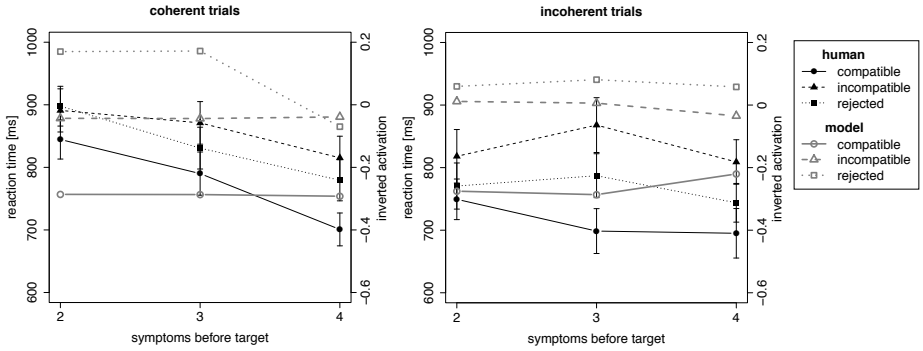


Figure 3.5 Inverted activation values from Simulation 1 and human reaction times for coherent (left) and incoherent trials (right). (Activation values are inverted so that they can be plotted directly against the reaction time data.)

Initial Activation of the Explanation Nodes

Simulation 1 and 2. One method to model sequential data in constraint-satisfaction models that might seem feasible is to use the output activation of the explanation nodes of one run as the input activation of these nodes in the next run (left side of Figure 3.1). Thus, activation values of explanation nodes are not reset before integrating a new symptom, but the values that resulted when integrating the previous symptom are used as start values. The observation of symptoms is modeled by connecting the currently observed symptom to the SEU (with a connection strength of .1). Subsequently, this model is referred to as Simulation 1.

The reason why this method does not work is the continuous influx of activation from the SEU through the currently observed symptom. Any activation at the beginning of a run is overwritten by the activation spreading from the SEU and only the connection strengths to the SEU determine the stable state of the network. This can be easily demonstrated by comparing the results of Simulation 1 to a model that is identical except for the fact that the explanation nodes are reset to zero after each run (Simulation 2). Simulation 1 and Simulation 2 produce basically the same activation results.

In Figure 3.5, the inverted activation values calculated by Simulation 1 are plotted against the human data for coherent (r = -.58) and incoherent trials (r = -.63). The model has an overall bad fit. Although compatible explanations are activated highest in the model as well as in the human data, the model does not show an increasing activation over the course of the trials as it is found in the human data. In incoherent trials, the models' activation even decreases with an increasing number of observed symptoms. Furthermore, contrary to the human data, rejected explanations in the model are activated less than incompatible explanations. Such a pattern of activation should only be expected if incoming information is not integrated properly.

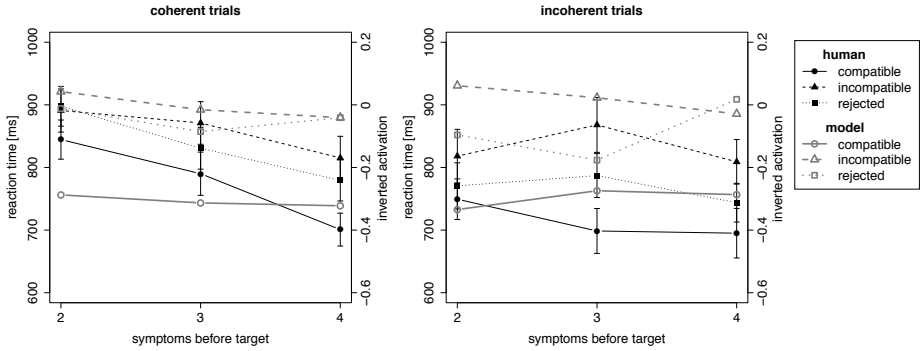


Figure 3.6 Inverted activation values from Simulation 3 and human reaction times for coherent (left) and incoherent trials (right).

Connection Strength to the SEU

Simulation 3. An alternative approach for modeling sequential data in constraint-satisfaction models is to use the connection strength between the evidence nodes and the SEU as proposed by Wang et al. (2006b) (see Figure 3.4 and right side of Figure 3.1). Contrary to Simulations 1 and 2, here not only the current symptom but also previously observed symptoms are connected to the SEU. The strength of the links to the SEU depends on the time elapsed since the respective symptom was observed. The most recently observed symptom gets a full connection to the SEU (.1). Earlier observations are connected to the SEU with a decayed strength as proposed by Wang et al. Before each run, the network is reset to its default values. That is, the activation of all chemicals and of all but the currently observed symptom is set to zero.

Again, the model was run for coherent (r = -.66) and for incoherent trials (r = -.73). As illustrated by Figure 3.6, this model produced a better fit than Simulations 1 and 2. As in the human data, compatible explanations receive the highest activation and incompatible explanations receive the lowest activation. However, the model again has difficulties to fit the change in activation over time. For example, in coherent trials, the model strongly underpredicts the increasing activation of compatible explanations over time.

Simulation 4. For better capturing the increasing activation over time, we presumed that the influence of each single symptom would need to be higher. Therefore, we developed a fourth model in which higher weights were given to the connection between observed symptoms and the SEU. The full connection, that is, the weighting of the most recent symptom, was now set to 1, and the respective decayed connections were calculated based on this value for the full connection. Except for this change, the model was identical to Simulation 3.

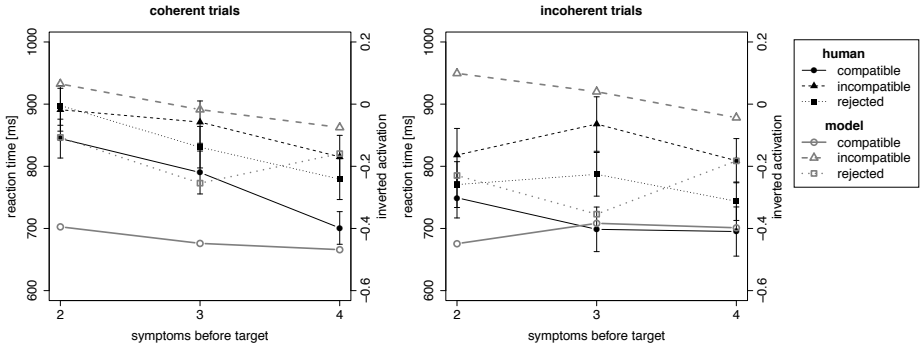


Figure 3.7 Inverted activation values from Simulation 4 and human reaction times for coherent (left) and incoherent trials (right).

Results of this model are shown in Figure 3.7 for coherent (r = -.70) and incoherent trials (r = -.81). As was to be expected, the differences between the different explanations increase compared to Simulation 3. Also the course of activation over time is fit better by this model. However, in coherent trials, the model still underpredicts the increase of activation over time for compatible explanations and it produces a pattern for rejected explanations that is not found in the human data. For incoherent trials, overall, the model predicts the difference between explanations, but it does not fit the change of activation over time.

In Simulations 3 and 4, the previous state of the models is retained by connecting not only the current, but also previous symptoms to the SEU. By letting the strength of these connections decay over time, the order of observed information is modeled. But is the decay of connection strengths necessary to model sequential information integration?

Simulation 5. To clarify this question, we developed a fifth model where, as in Simulations 3 and 4, all previously observed symptoms are connected to the SEU. However, previous symptoms do not decay, but they keep the full connection strength of 1.

Results are shown in Figure 3.8 for coherent (r = -.75) and incoherent trials (r = -.85). For coherent trials, this simplified version of the model produces the best fit to the human data. It shows the activation differences between the explanations and it fits the activation pattern over time considerably well. However, also this model has difficulties in fitting the incoherent trials. Whereas the participants' reaction times reflect a change in their diagnosis in the light of the new, incoherent evidence (the third symptom of the trial), the model does not produce a clear activation difference between compatible and rejected explanations after the incoherent evidence is observed.

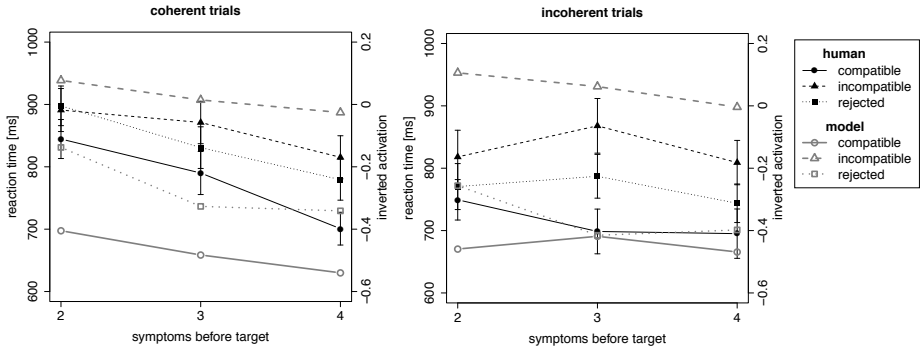


Figure 3.8 Inverted activation values from Simulation 5 and human reaction times for coherent (left) and incoherent trials (right).

Conclusion

We evaluated two possible approaches for modeling sequential information integration in diagnostic reasoning. These approaches differed in the mechanism implemented to integrate new information with information obtained earlier. In the first approach, activation results from the previous integration step were carried over to the next step, where they were integrated with the new information. In the second approach, the previous state of the network was preserved more indirectly, by connecting not only the current but also earlier observations to the "energy source" (the SEU) of the network.

Results show that the first approach (Simulations 1 and 2) does not work. The initial activation of the network's nodes is overwritten by the activation spreading from the SEU. The second approach was more successful. Following a suggestion from Wang et al. (2006b), we implemented versions of models that differed with respect to how strongly symptoms that were observed over time influence the current activation of the network (Simulations 3 and 4). Both models were able to reproduce the activation differences between explanations found in the human data, with higher weights of the connection between SEU and observed symptoms resulting in better model fits. However, also these models had difficulties in fitting the course of activation over time. A simplified version of these models (Simulation 5), where the influence of earlier evidence did not decay over time, produced a surprisingly high fit in coherent trials, but failed to model the time course of activation in incoherent trials.

Concluding, our results support the approach for modeling sequential information integration as it was proposed by Wang et al. (2006b). However, our results suggest the parameter setting proposed by Wang et al. to be reconsidered. To model the course of activation during the task, we needed to implement a much higher amount of activation spreading from the observed symptoms. Furthermore, our results suggest

that not yet explained observations do not decay over time, as suggested by Wang et al., but retain a stable influence on the network.

We must stress that none of the models was able to sufficiently fit the pattern of activation in incoherent trials. Although Simulations 3 and 4, where observations decayed over time, produced at least the differences between explanations as found in the human data, they did not model the course of activation adequately. This might have several reasons. First, the implementation of constraint satisfaction may be inappropriate. Second, and more plausible given the success of constraint satisfaction models in various areas, the deviation between human and model data demonstrates the involvement of more conscious reasoning processes during incoherent trials. In coherent trials, the automatic activation processes modeled by the constraint satisfaction networks is perfectly sufficient to solve the task. In incoherent trials however, a pure activation-based approach struggles. Nodes would have to be added or connections other than connections to the SEU would have to be manipulated. As we discussed in more detail in Chapter 2, to fully capture cognitive processes involved in such trials and in tasks with more complex knowledge structures, hybrid modeling approaches which allow for investigating the interaction of automatic memory activation with more deliberate reasoning strategies, might be promising.

The Influence of Experience and Context on Hypothesis Generation

In which I use behavioral experiments and an ACT-R model to investigate the respective contribution of previous experience and the current context on explanations' availability in memory.

Chapter

The Influence of Experience and Context on Hypothesis Generation

Chapter

Abstract

Recent theories of diagnostic reasoning propose that automatic memory activation processes are involved in the generation of hypotheses from memory. Two aspects have been suggested to play a role: (1) a hypothesis' past usefulness and (2) its usefulness in the current context. Based on a general theory of memory, we present two mechanisms that might explain theses aspects: (1) a hypothesis' base-level activation, reflecting its past usefulness, and (2) the spreading activation it receives from current observations, reflecting its usefulness in the current context. We conducted an experiment in which participants had to generate hypotheses, while both memory components were independently manipulated by an ostensibly unrelated secondary task. The results show an effect of both manipulations and are quantitatively predicted by an ACT-R model in which we implemented both memory mechanisms. Discrepancies between the behavioral data and the predictions of the mere memory-based model were also revealed and their potential reasons are discussed.

Introduction

A doctor trying to find the best diagnosis for his patient's symptoms, a scientist trying to understand her data, or a person trying to deduce someone else's intentions are all examples in which hypotheses need to be generated and evaluated from memory. Traditionally, cognitive psychology has focused on deliberate reasoning processes that are engaged in solving such tasks (e.g., T. R. Johnson & Krems, 2001; Klahr & Dunbar, 1988). More recently, researchers have started investigating the role of automatic memory processes for hypothesis generation (e.g., Arocha & Patel, 1995; Mehlhorn, Taatgen, Lebiere, & Krems, 2011; Thomas et al., 2008; see also Chapter 2 of this thesis).

The basic idea is that automatic memory processes can provide an adaptive subset of possible hypotheses from memory, which can serve as input to a more deliberate evaluation process (Thomas et al., 2008). Such a distinction between a memory-based and a reasoning-based component is also a central aspect of dual-process theories. They assume fast, automatic processes to provide a possible answer, which might then be justified or revised by more time consuming, deliberate reasoning (Evans, 2008).

Empirical evidence suggests that the generation of hypotheses from memory depends on two aspects. A first aspect is the hypotheses' usefulness in the past. It has been shown that from all potential hypotheses, reasoners tend to generate those that have a high a priori probability based on previous experiences (Dougherty & Hunter, 2003a; Weber et al., 1993). A second aspect is the hypotheses' usefulness in the current context. For example, Weber et al. (1993) have shown that physicians generate those diagnoses that are most likely in the light of a patient's symptoms. While both aspects have received empirical support, the underlying mechanisms, as well as their respective contribution for hypothesis generation have received relatively little attention in the literature (for an exception, see Thomas et al., 2008).

The goal of this chapter is to show the respective contribution of both aspects and to test in how far general memory mechanisms can explain the effects. We first give an overview on the memory mechanisms, before we describe an experiment in which we manipulated both aspects. Subsequently, we show how we generated quantitative predictions from a cognitive model and compare these predictions to the data from the experiment.

Memory Processes in Hypothesis Generation

A general assumption of memory theories is that "the memory system […] makes most available those memories most likely to be needed" (Anderson, 2007, p. 109). How does it do that? While theories differ on the exact proposed mechanisms and the used vocabulary, commonly, two components are assumed to determine the likelihood of an item to be needed from memory: its a priori probability based on previous experiences and its usefulness in the current context.

Previous experience. Anderson and Schooler (1991) investigated how previous experience predicts an item's likelihood to be needed from memory. Based on their

results it has been suggested (e.g., Anderson, 2007), that the inherent availability of an item in memory can be described by its base-level activation, B_i, which depends on the frequency and recency of the items past usage:

$$B_i = \ln\left(\sum_{k=1}^{n} t_k^{-d}\right) \tag{4.1}$$

where n is the number of previous encounters with item i, t_k is the time since the k^{th} encounter, and d is a decay parameter (producing the power law of forgetting). Using the example of a physician, this mechanism could for example explain why, especially during flu season, the flu will seem to be a more likely diagnosis for a patient's symptoms than throat cancer.

Current context. Various memory theories share the assumption that information in the environment can serve as a cue for the retrieval of items from memory (e.g., Anderson, 2007; Kintsch, 1998; Thomas et al., 2008). A frequently proposed mechanism underlying this cued retrieval is spreading activation between observed information and associated items in memory (e.g., Anderson, 2007; Thagard, 2000). Specifically, Anderson proposes that an item i in memory receives spreading activation, S_i, from each associated piece of information j, which is currently stored in working memory:

$$S_i = \sum_j W_j S_{ji} \tag{4.2}$$

where W_j represents a weighting of j in working memory and S_{ji} represents the associative strength between i and j. This associative strength reflects the extent to which an observed piece of information increases the likelihood of an associated item to be needed from memory. Using the physician's example again, this mechanism could explain why, in the context of symptoms that point specifically at throat cancer, this diagnosis might become available in memory.

In a previous study (Mehlhorn et al., 2011; see also Chapter 2 of this thesis), we investigated whether the current context could indeed affect the availability of hypotheses in memory as predicted by such a spreading activation account. In two experiments, participants had to generate diagnoses for sequentially presented medical symptoms, while we tracked the availability of different hypotheses in memory with a probe reaction task. As predicted, availability was found to vary over time as a function of the observed symptoms.

Respective contribution of the memory processes. While the results of Mehlhorn et al. (2011) provide evidence for the influence of the current context via spreading activation mechanisms, the respective contribution of a hypothesis' past usefulness was not investigated in that study. Anderson (2007) argued that base-level activation, B_i, and spreading activation, S_i, are independent additive components that determine the availability of an item i in memory:

$$A_i = B_i + S_i + \varepsilon \tag{4.3}$$

where A_i is an item's activation in memory and ε is a random noise component (see e.g., Anderson, 1990, for the underlying Bayesian statistics). For our physician, this could, explain why, depending on whether the base-level or the spreading-activation component are stronger, the flue or throat cancer are more strongly available in memory.

Overview of the Experiment

To investigate the memory components outlined above, we conducted an experiment in which participants had to solve a *diagnostic reasoning task,* while at the same time carrying out a secondary *choice-reaction task*. In each experimental trial, the medical symptoms of a hypothetical patient were presented one at a time on the screen. At the end of the trial, participants had to report the diagnosis that explained the patient's symptoms. During presentation of the symptoms, participants were auditorily presented with letters and had to indicate as fast as possible whether the letter was a consonant (target) or a vowel. This choice-reaction task was used to manipulate both memory components independently.

The base-level component was manipulated by the *targets* presented in the choice-reaction task. Targets were either neutral consonants or consonants that were also used to name potential diagnoses. We expected that retrieving such diagnosis-naming targets from memory would increase the base-level activation of the respective diagnoses. Consequently, performance in the diagnosis task should be reduced, as the diagnoses whose base-level activation was increased by the diagnosis-naming targets are not necessarily the correct diagnoses for the presented symptoms.

The spreading activation component was manipulated by requiring participants to *count* the targets presented in the choice-reaction task in part of the trials. The idea behind this manipulation is that counting, as well as generating hypotheses, both rely on a central working-memory resource, which can only be used for one task at a time (Borst et al., 2010; Oberauer, 2002). The to be expected working memory conflicts can result in losing part of the observed symptoms from working memory. Consequently, performance in the diagnosis task should be reduced, as the correct diagnosis receives less spreading activation from the reduced amount of symptoms in working memory.

Method

Participants

Twenty-five native German speaking undergraduate students from the University of Groningen took part in this experiment for course credit (19 female; mean age 21.2, *SD* = 1.3).

Material

Diagnostic knowledge. The knowledge participants needed to learn before solving the diagnosis task was adapted from Mehlhorn et al. (2011; see also Chapter 2 of

Table 4.1 Overview of diagnostic knowledge participants had to acquire before the experiment (original material in German).

Aggregate state and source of contamination	Chemical	Specific symptoms			Unspecific symptoms		
Gasiform, inhaled	B	Cough		Shortness of breath	Headache		Dizziness
	T	Cough	Sneezing		Headache	Fever	
	W		Sneezing	Shortness of breath		Fever	Dizziness
Crystalline, skin contact	L	Redness		Rash	Headache		Dizziness
	H	Redness	Itching		Headache	Fever	
	G		Itching	Rash		Fever	Dizziness
Liquid, drinking water	C	Diarrhea		Cramps	Headache		Dizziness
	M	Diarrhea	Vomiting		Headache	Fever	
	R		Vomiting	Cramps		Fever	Dizziness

this thesis) and consisted of nine hypothetical chemicals (all single consonants, see Table 4.1). The chemicals were grouped into three artificial categories and caused four symptoms each. To reflect the complexity of real-world diagnostic knowledge, symptoms were either specific for a category (e.g., cough) or unspecific (e.g., headache).

Audio stimuli. For the choice-reaction task we generated three sets of audio files: *chemical consonants* (the 9 chemical names), *non-chemical consonants* (the letters SKQFVDNP), and *vowels* (the letters AEIOUÄÖÜ). In the *non-chemical condition,* audio stimuli were randomly sampled from the non-chemical consonants and vowels. In the *chemical condition*, audio stimuli were randomly sampled from the chemical consonants and vowels. In this condition, we additionally varied whether the set of consonants included the correct diagnosis for the presented symptoms (*correct diagnosis primed*) or not (*correct diagnosis not primed*).

Procedure

Training. To learn the diagnostic knowledge, participants were visually presented with the four symptoms caused by one of the chemicals and had to enter a diagnosis. By receiving feedback, they learned which chemicals where associated with which symptoms. Categories were first trained separately, before the same training was repeated for the complete material. The order of categories, diagnoses, and symptoms

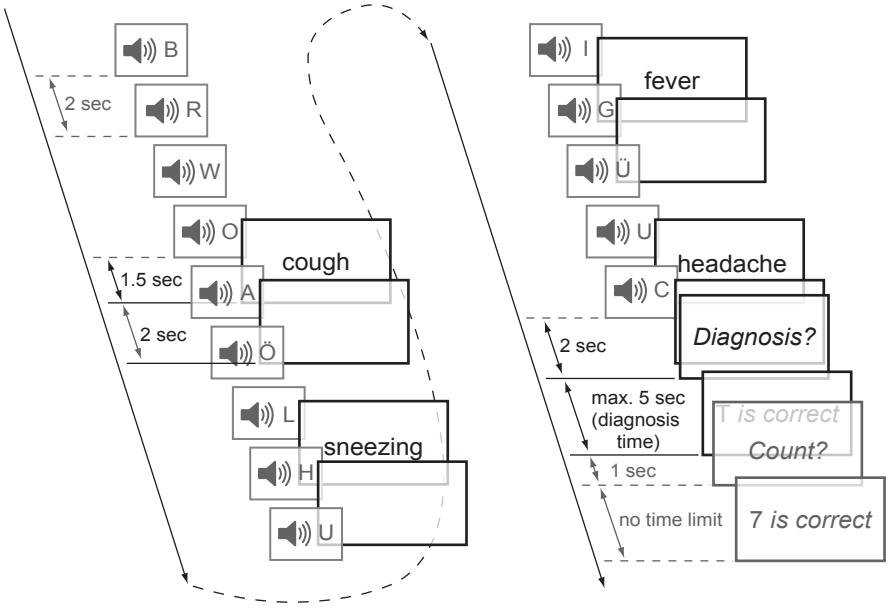


Figure 4.1 Sample trial for the chemical–count condition. In this trial, 14 audio stimuli were presented of which 7 were consonants. T was the correct diagnosis for the medical symptoms.

was randomized for each participant. After acquiring the diagnostic knowledge (performance criterion: 100%), the choice-reaction task was first practiced alone and then in combination with the diagnosis task. Participants were informed that in the experiment it was important to do both tasks as fast and accurately as possible.

Experiment. The experiment was split into 8 blocks. In each block, 9 trials from one of the four conditions (non-chemical – no count, non-chemical – count, chemical – no count; chemical – count) were presented. The conditions were assigned to the blocks in random order, with the constraint that each condition had to be presented once in the first four and once in the second four blocks.

In each trial, 12 to 14 audio stimuli were presented with a SOA of 2 s (Figure 4.1). Six to twelve of the stimuli were consonants (the exact numbers were randomly drawn from a uniform distribution for each participant for each trial). Additionally, the four symptoms of one of the chemicals were visually presented for 2 s each. The 1st symptom was presented 1.5 s after the onset of the 2nd, 3rd, or 4th audio stimulus, with the exact position depending on the total number of stimuli in the trial. Before each subsequent symptom, 3 audio stimuli were presented. The final audio stimulus was presented .5 s after onset of the 4th symptom. The presentation order of audio stimuli and symptoms was randomized for trials and participants. Each chemical occurred with equal frequency as correct diagnosis in each block. After all stimuli had been presented, participants had to enter their diagnosis within maximum 5 s and, in the count condition, to enter the number of consonants. Participants received

visual feedback for their diagnosis and count. Auditory feedback was presented for the choice reactions if the response was wrong or not given within 1.25 s.

Model

Based on a previously published model of hypothesis generation (Mehlhorn et al., 2011; see also Chapter 2 of this thesis), we implemented a cognitive model in the ACT-R architecture (Anderson, 2007). ACT-R makes precise predictions about how the memory mechanisms described above affect the probability and latency of memory retrieval. It allows for modeling the task, as solved by the participant, and thereby produces results that are directly comparable to the human data. Below we briefly describe the model (the model code, including more detailed explanations, can be downloaded from http://www.ai.rug.nl/~katja/models).[1]

The model is presented with the same tasks as the participants, that is, it has to discriminate between the auditorily perceived consonants and vowels, it has to count the consonants (in the count condition), and it has to generate a diagnosis for the visually perceived symptoms. The knowledge necessary to solve these tasks is represented in the model's long-term memory (the declarative memory).

To solve the choice-reaction task, the model tries to retrieve the perceived letter from declarative memory, assesses if the retrieved letter is a consonant or vowel, and enters its response. To count, the model keeps track of the current count in working memory.[2] The count is incremented when a retrieved letter is classified as consonant. When asked for the count, the model enters the current count. To solve the diagnosis task, the model stores the observed symptoms in working memory, from where they spread activation to associated diagnoses in declarative memory. When asked for the diagnosis, the model retrieves and enters that diagnosis from declarative memory that has the highest activation as calculated by Equation 4.3.

In the no-count condition, all observed symptoms are stored in working memory until the model is asked for a diagnosis. In the count condition, the set of symptoms that is currently stored in working memory has to be swapped out to declarative memory whenever the model needs working memory for counting, because working memory can only be used for one task at a time (see Borst et al., 2010, for empirical support). Whenever working memory is needed for the diagnosis task again, the current count is swapped out and the set of symptoms is swapped back in. Information can be lost during swapping because, due to noise, the model might erroneously retrieve older working-memory contents (e.g., an incomplete set of symptoms) from declarative memory.

[1] *To fit the model, we estimated the latency and stochasticity of memory retrievals, the base-level activation of facts in memory at the beginning of each trial, and the maximum associative strength between items. Based on earlier results (Mehlhorn et al., 2011), we assumed the associative strength of each symptom in working memory to be weighted by a constant value of W, independent of the number of symptoms (see the model code for the exact parameter values). The model was run 40 times for each participant. Results were calculated for each run and then averaged across runs.*

[2] *To model working memory we use one of the buffers of ACT-R's cognitive modules, the imaginal buffer. The imaginal buffer is commonly used to hold a mental representation of the problem currently in the focus of attention (Borst et al., 2010).*

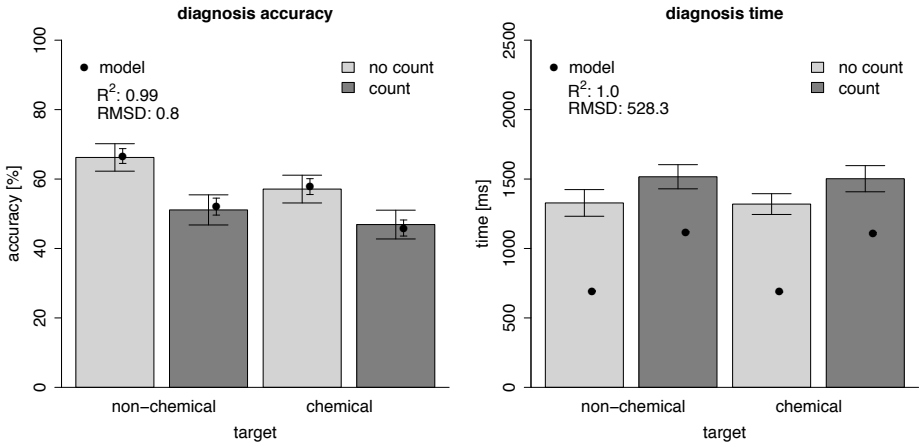


Figure 4.2 Diagnosis accuracy (left) and time (right) for the factors target (non-chemical, chemical) and counting (no count, count); $M \pm 1$ *SE* for human (bars) and model data (dots).

Results

Performance in the Diagnosis Task

Effects of target and counting. To investigate the respective impact of both manipulated memory components, we analyzed diagnosis accuracy and time for the factors target (non-chemical, chemical) and counting (no count, count). The results are shown in Figure 4.2 and Table 4.2. As correctly predicted by the model, chemical targets lead to lower diagnosis accuracy than non-chemical targets, but do not affect diagnosis times. Also the effects of counting are correctly predicted by the model: counting leads to lower diagnosis accuracies and higher diagnosis times than no counting. However, the model generally underestimates diagnosis times.

Effect of priming the correct diagnosis. To further test the effect of the base-level manipulation on diagnosis performance, we compared chemical-condition trials in which the correct diagnosis was among the presented chemical consonants (*correct diagnosis primed*) to chemical-condition trials in which this was not the case (*correct diagnosis not primed*). As shown in Table 4.3, the model predicts higher diagnosis accuracy in primed than in not-primed trials, while participants do not show this effect on diagnosis accuracy. However, participants do show an effect on diagnosis time, with primed diagnoses being faster than not-primed ones. The model correctly predicts this effect on diagnosis time, but underpredicts its magnitude.

Performance in the Choice-Reaction Task

Reaction and counting accuracy. Participants discriminated between consonants and vowels with a reasonably high accuracy (M = 88.5%, SD = 6.0), which is approximated

Table 4.2 Results of repeated-measures ANOVAs for the factors target (non-chemical, chemical) and counting (no count, count).

Dependent measure	Effects	*F*	*p*	η_p^2
Diagnosis accuracy	Main effect of target (non-chemical, chemical)	(1,24) = 11.15	**.003**	.317
	Main effect of counting (no count, count)	(1,24) = 27.47	< **.001**	.534
	Interaction of target and counting	(1,24) = 2.60	.120	.098
Diagnosis time[a]	Main effect of target (non-chemical, chemical)	(1,23) = .07	.792	.003
	Main effect of counting (no count, count)	(1,23) = 14.72	< **.001**	.390
	Interaction of target and counting	(1,23) = .01	.942	< .001

Note. *p* values <.1 are shown in bold. [a] Only correct responses. Data points that differed more than 3 *SD* from a participant's condition mean were excluded (0.1% of the diagnosis time data). Additionally, one participant had to be completely excluded from this analysis, because of a diagnosis accuracy of 0 in the non-chemical – count condition.

Table 4.3 Human and model data in the chemical-consonants condition, depending on whether the correct diagnosis was primed or not.

Dependent measure	Priming	Human *M* (*SD*)	Model *M* (*SD*)
Diagnosis accuracy [%]	correct diagnosis primed	52.1 (22.5)	58.4 (11.4)
	correct diagnosis not primed	51.7 (22.8)	45.1 (11.7)
		$t(24) = .08, p = .940$	
Diagnosis time [ms][a]	correct diagnosis primed	1339 (363)	853 (53)
	correct diagnosis not primed	1533 (551)	902 (66)
		$t(24) = -2.50, p =$ **.020**	

Note. *p* values <.1 are shown in bold. For simplicity, here we collapsed over the factor counting, which was justified by a lack of interactions with the factor. [a] Only correct responses. No diagnosis times differed more than 3 *SD* from a participant's condition mean.

well by the model (*M* = 91.3%, *SD* = 1.0). The correct count was reported in 57.1% (*SD* = 20.2) of the trials in the count condition. The model reaches a slightly lower counting accuracy (*M* = 39.7%, *SD* = 8.3).

Reaction accuracy over the trial. Due to the nature of memory activation, we also expected the diagnosis task to influence performance in the choice-reaction task. For

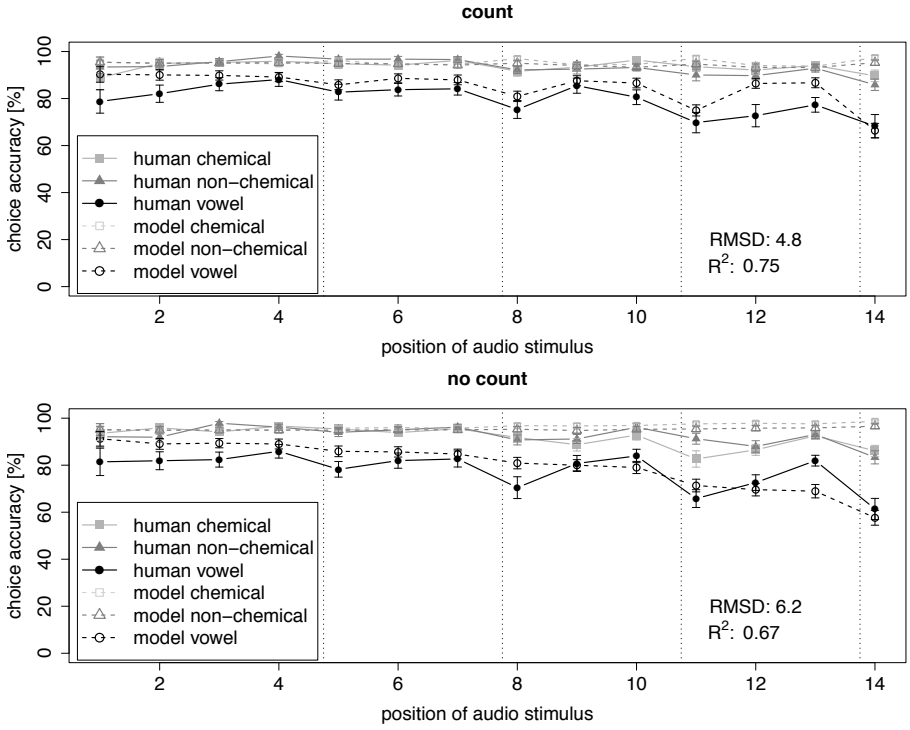


Figure 4.3 Human and model accuracy ($M \pm 1$ *SE*) for the choice-reaction task in the count (top) and no-count (bottom) condition for the different stimuli (non-chemical, chemical, vowel) at each of the 14 possible positions in the trial (positions were lined out to the last stimulus, so that each trial had a 14th stimulus, but only trials with 14 stimuli had a 1st stimulus). Vertical lines indicate the four points at which symptoms were presented.

example, if observed symptoms indeed spread activation to associated hypotheses, the availability of the respective chemical consonant letters should increase, resulting in an increased chance to retrieve consonants and thereby to more errors in reacting to vowels. To test this assumption, we analyzed the accuracy of reactions to the different types of stimuli (chemical consonant, non-chemical consonant, vowel), depending on counting, and on the stimulus' position in the trial (1-14).

As correctly predicted by the model, overall accuracy in the choice-reaction task was indeed lower for vowels than for consonants (Figure 4.3). In the count condition, the model correctly predicts a drop of response accuracy to vowels whenever a new symptom is observed. This happens because, due to the swapping of information between the diagnosis and counting task, observed symptoms spread activation to associated chemical consonants at these points. In the no-count condition, the model predicts a slightly more gradual decrease of response accuracy to vowels than found in the human data. The decrease in the model is caused by the increasing amount of spreading activation from the increasing number of symptoms in working memory.

Discussion

Empirical research has shown that, when generating hypotheses from memory, reasoners generate only a small subset of all potential hypotheses. However, this subset seems to be highly adaptive, as it contains those hypotheses that have (1) a high a priori probability based on previous experience and (2) a high usefulness in the current context. The results of our study can help to understand this adaptive selection in terms of general memory mechanisms. We presented base-level activation as a memory mechanism that is sensitive to a hypothesis' past usefulness. It predicts the availability of a hypothesis in memory to increase with the frequency and recency of its usage. We presented spreading activation as a mechanism that can regulate the influence of the current context on a hypothesis' availability in memory. It predicts the availability of a hypothesis in memory to increase with the amount of observations in working memory that are associated with this hypothesis and with the strength of their association.

While the influence of the current context via spreading activation mechanisms was already supported in an earlier study (Mehlhorn et al., 2011; see also Chapter 2 of this thesis), the respective contribution of a base-level activation mechanism that reflects a hypothesis' past usefulness had not yet been shown. To test this contribution, we manipulated both components within one experiment. Diagnosis performance showed main effects of *both* manipulations, suggesting that the components might indeed reflect two distinct aspects of memory activation. This assumption is supported by the cognitive model, which revealed both base-level activation and spreading activation to be important components for fitting the behavioral data.

The model explains the reduced performance in the chemical compared to the non-chemical condition by an increase in base-level activation of wrong diagnoses, which were presented as letters in the choice-reaction task. Finding behavioral evidence for this influence is especially interesting because, objectively, participants had to do the same choice-reaction task in both conditions: discriminate between consonants and vowels. It is additionally interesting because the letters were used in semantically different meanings in both tasks: In the choice-reaction task, they were used to discriminate between consonants and vowels, while in the diagnosis task, they were used as potential diagnoses. Nevertheless, diagnosis performance decreased, as predicted by the model, when the consonants of the choice-reaction task were names of chemicals. This demonstrates the impact of automatic memory activation on hypothesis generation.

The model explains the reduced performance in the count condition compared to the no-count condition by a decrease in spreading activation to the correct diagnosis. This decrease is caused by working-memory conflicts between the count and diagnosis task, which result in the loss of observed symptoms from working memory in the count condition. Together with the findings presented in Chapter 2, this illustrates how hypothesis generation depends on the current context that is available in working memory.

It has been proposed that automatic hypothesis-generation processes interact with deliberate hypothesis evaluation (Thomas et al., 2008). Deviations between the results of our merely memory-based model and the behavioral data suggest such an interaction also in our study. The absence of an effect of priming on human diagnosis accuracy, as well as the model's general underprediction of diagnosis times, suggest that participants did not simply enter the diagnosis made most available by memory activation as in the model. Rather, participants might have used additional time to evaluate and justify the retrieved hypotheses.

De Neys (2006) showed that the use of deliberate reasoning strategies increases with the availability of working-memory resources. Such an increased use of deliberate reasoning might explain why, in our choice-reaction data, the model which did not use any deliberate reasoning strategies fitted better in the count condition (with high working memory demands) than in the no-count condition (with lower working memory demands). However, despite participants' potential additional use of deliberate reasoning, the mere activation-based model fits the choice-reaction data quite well. This is remarkable, because this fit directly emerges from the memory activation mechanisms implemented in the model, without us adding any additional assumptions.

Are the results of our laboratory study generalizable? Real-world hypothesis generation will often be more complex and less structured than the diagnosis task participants solved in our experiment. We decided for such a simplified hypothesis-generation task, because it allowed for the experimental control necessary to test our assumptions about the subtle effects of memory activation. However, we do not expect this simplification of the task to limit the validity of our results. Higher complexity and a less well defined task structure are expected to even increase the importance of memory activation processes (Dijksterhuis & Nordgren, 2006).

Before closing, we want to stress that the main point of this chapter is not to promote one particular memory theory, but to show how taking into account the importance of automatic memory activation can help to understand hypothesis generation. While memory theories differ in the exact proposed mechanisms, many theories share the assumption that the probability of an item to be needed from memory depends on the two factors discussed in this chapter: the item's a priori probability based on previous experiences and its usefulness in the current context (e.g., Anderson, 2007; Thomas et al., 2008).

Thirty-Nine ACT-R Models of Decision Making

In which I use ACT-R to compare different strategies that have been proposed for how people make decisions based on the availability of information in memory.

This chapter is accompanied by Supplementary Online Materials which can be found at http://www.ai.rug.nl/~katja/thesis

An earlier version of this chapter was published as:

Marewski, J. N. & Mehlhorn, K.* (2011). Using the ACT-R architecture to specify 39 quantitative process models of decision making. *Judgment and Decision Making, 6,* 439-519.

**alphabetical order, authors contributed equally*

Thirty-Nine ACT-R Models of Decision Making

Abstract

Hypotheses about decision processes are often formulated qualitatively and remain silent about the interplay of decision, memory, and other cognitive processes. At the same time, existing decision models are specified at varying levels of detail, making it difficult to compare them. We provide a methodological primer on how detailed cognitive architectures such as ACT-R allow remedying these problems. To make our point, we address a controversy, namely, whether noncompensatory or compensatory processes better describe how people make decisions from the accessibility of memories. We specify 39 models of accessibility-based decision processes in ACT-R, including the noncompensatory recognition heuristic and various other popular noncompensatory and compensatory decision models. Additionally, to illustrate how such models can be tested, we conduct a model comparison, fitting the models to one experiment and letting them generalize to another. Behavioral data are best accounted for by race models. These race models embody the noncompensatory recognition heuristic and compensatory models as a race between competing processes, dissolving the dichotomy between existing decision models.

Introduction

Even if the mind has parts, modules, components, or whatever, they all mesh together to produce behavior. [...] If a theory covers only one part or component, it flirts with trouble from the start. (Allen Newell, 1990, p. 17)

One way to increase the precision of theories of decision making is to specify the cognitive processes decision-making mechanisms are assumed to draw on. Corresponding *process models* predict not only what decision a person will make, but also how the information used to make the decision will be processed. The past decades have seen repeated calls to develop process models, and in fact, such models have become increasingly popular (e.g., Brandstätter, Gigerenzer, & Hertwig, 2006; Einhorn, Kleinmuntz, & Kleinmuntz, 1979; Ford, Schmitt, Schechtman, Hults, & Doherty, 1989; Gigerenzer & Goldstein, 1996; Gigerenzer et al., 1991; Marewski, Gaissmaier, & Gigerenzer, 2010a, 2010b; Payne, Bettman, & Johnson, 1988, 1993; Schulte-Mecklenbeck, Kühberger, & Ranyard, 2010). The predictions made by these models have motivated a number of debates; for example, whether people rely on noncompensatory, lexicographic as opposed to compensatory, weighted-additive processes in inference, choice, and estimation (e.g., Bergert & Nosofsky, 2007; Bröder & Schiffer, 2003, 2006; Cokely & Kelley, 2009; E. J. Johnson, Schulte-Mecklenbeck, & Willemsen, 2008; Lee & Cummins, 2004; Marewski, 2010; Mata, Schooler, & Rieskamp, 2007; B. R. Newell, Weston, & Shanks, 2003; Nosofsky & Bergert, 2007; Rieskamp & Hoffrage, 1999, 2008; Rieskamp & Otto, 2006; von Helversen & Rieskamp, 2008).

Yet, often such process models are underspecified relative to the process data against which they can be tested. In this chapter, we show how precision can be lent to process models by implementing them in a cognitive architecture. We will make our point by focusing on a class of models that assume people to make decisions by exploiting the *accessibility* (e.g., Bruner, 1957; Higgins, 1996; Kahneman, 2003) of memory contents. These models have been at the focus of a debate about what processes describe people's decisions best when they make inferences about unknown states of the world; such as when predicting which sports teams are likely to win a competition, which politician will win an election, or which cities are likely to grow fastest in the number of inhabitants.

A Case Study of Underspecified Process Hypotheses

Numerous accessibility-based decision models have been proposed, featuring concepts such as familiarity, fluency, availability, or recognition (e.g., Dougherty, Gettys, & Ogden, 1999; Jacoby & Dallas, 1981; Koriat, 1993; Pleskac, 2007; Tversky & Kahneman, 1973). One such model is the *recognition heuristic* (Goldstein & Gigerenzer, 2002). As suggested by its name, this simple decision strategy operates on our ability

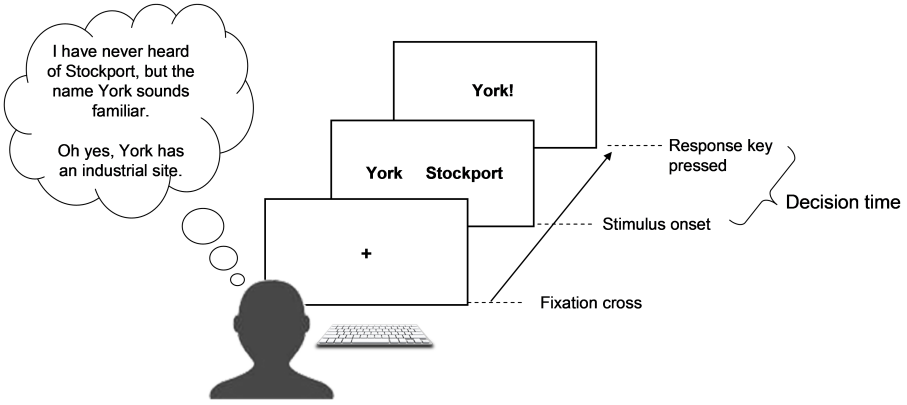


Figure 5.1 The memory paradigm. In a two-alternative forced-choice task, on a computer screen a person is first shown a fixation cross, and thereafter presented with the names of two alternatives (e.g., two city names). The person's task is to infer which of the two has a larger value on the criterion (e.g., which of two cities is larger). To make this decision, the person has to retrieve all information she wants to use from memory. For instance, the person may believe to recognize a city's name and additionally remember that the city has an industrial site, suggesting that it is a large city. Once a person has made her decision, she presses a key to respond. Gigerenzer and Goldstein (1996) referred to such experimental paradigms as inferences from memory.

to discriminate between *recognized* alternatives that we have encountered in our environment before, and *unrecognized* ones that we do not remember to have seen or heard of before. In doing so, the heuristic can help us to infer which of two alternatives (e.g., two cities, York and Stockport), one recognized and the other not, has the larger value on an unknown *criterion* (e.g., city size). The heuristic reads as follows: *If only one of two alternatives is recognized, infer the recognized one to be larger.*

The recognition heuristic is a noncompensatory model for memory-based decisions: Even if further knowledge beyond recognizing an alternative is retrieved, this knowledge is ignored when the heuristic is used. Instead, the decision is based solely on recognition. In contrast to the recognition heuristic and related accessibility-based heuristics (e.g., Schooler & Hertwig, 2005), many other decision models posit that people evaluate alternatives by using knowledge about their attributes as *cues* (Bröder & Schiffer, 2003; Hauser & Wernerfelt, 1990; Lee & Cummins, 2004; Payne et al., 1993). For instance, to infer which of two cities is larger, a person could rely on one of the classic compensatory unit-weight linear *integration strategies* (e.g., Dawes, 1979): The person could recall whether the cities have industry sites, airports, or famous soccer teams. For each city, the person could count the number of *positive* and *negative* cues (e.g., having an airport would be a positive cue and lacking one a negative cue) and then infer the city with the larger sum to be larger (Einhorn & Hogarth, 1975; Gigerenzer & Goldstein, 1996; Huber, 1989). The assumption in such compensatory models is that an alternative's value on one cue is traded off against its value on another cue.

Process Hypotheses in the Memory Paradigm

The recognition heuristic has triggered a debate about what processes describe people's decisions best when they make inferences from the accessibility of memories: Do people rely on this noncompensatory heuristic, ignoring further knowledge, or do they use compensatory strategies instead? (Bröder & Eichler, 2006; Davis-Stober, Dana, & Budescu, 2010; Dougherty, Franco-Watkins, & Thomas, 2008; Erdfelder, Küpper-Tetzel, & Mattern, 2011; Gaissmaier & Marewski, 2011; Gigerenzer & Brighton, 2009; Gigerenzer & Goldstein, 2011; Gigerenzer, Hoffrage, & Goldstein, 2008; Glöckner & Bröder, 2011; Goldstein & Gigerenzer, 2011; Hertwig, Herzog, Schooler, & Reimer, 2008; Hilbig, Erdfelder, & Pohl, 2010; Hilbig & Pohl, 2009; Hochman, Ayal, & Glöckner, 2010; Hoffrage, 2011; Marewski, Gaissmaier, Schooler, Goldstein, & Gigerenzer, 2009, 2010; Marewski, Pohl, & Vitouch, 2010, 2011a, 2011b; McCloy, Beaman, & Smith, 2008; B. R. Newell & Shanks, 2004; Oeusoonthornwattana & Shanks, 2010; Oppenheimer, 2003; Pachur, 2010, 2011; Pachur & Biele, 2007; Pachur & Hertwig, 2006; Pachur, Mata, & Schooler, 2009; Pachur, Todd, Gigerenzer, Schooler, & Goldstein, 2011; Pohl, 2006, 2011; Reimer & Katsikopoulos, 2004; Richter & Späth, 2006; Scheibehenne & Bröder, 2007; Volz et al., 2006).

In this debate, many researchers have used the *memory paradigm* shown in Figure 5.1. The time it takes a person to make the decision – the *decision time* measured from stimulus onset until the person presses a key – is used to test hypotheses about the processes underlying the decision (Hertwig et al., 2008; Hilbig & Pohl, 2009; Marewski, Gaissmaier, Schooler et al., 2010; Richter & Späth, 2006; Volz et al., 2006). For instance, Pachur and Hertwig (2006) hypothesized that recognition memory would be more easily assessed than memories about cues, enabling people to make decisions based on the recognition heuristic faster than decisions based on cues.

Importantly, although tests of such process hypotheses are central to the debate about the recognition heuristic, thus far the hypotheses put forward in this debate lack precision. First, in the memory paradigm, in no study were decision times actually quantitatively predicted. Rather, mostly qualitative (e.g., ordinal) decision time hypotheses were tested. Second, in no study these hypotheses took into account the interplay among perceptual, memory, decision, intentional, and motor processes governing decision times in the memory paradigm (but see Marewski, 2008; Marewski & Schooler, 2011). In a recent test of process hypotheses with the memory paradigm, Hilbig and Pohl (2009), for example, derived qualitative decision time hypotheses for the recognition heuristic and compared them against corresponding hypotheses they derived from *evidence accumulation* processes, as they have been outlined by B. R. Newell (2005) and others (e.g., Lee & Cummins, 2004). Broadly speaking, the assumption of such evidence accumulation processes is that evidence (e.g., cues and other information) for each of two alternatives is accumulated sequentially until a decision threshold is reached (e.g., D cues are retrieved) and a decision made (e.g., in favor of the alternative with most accumulated evidence). In testing their hypotheses, Hilbig and Pohl subsumed a number of models under this broad notion of evidence accumulation, including a connectionist *parallel constraint satisfaction model* (Glöckner

& Betsch, 2008), and *decision field theory* (Busemeyer & Townsend, 1993). According to them, their decision time data could be accounted for by compensatory evidence accumulation models but were inconsistent with the recognition heuristic. However, Hilbig and Pohl did not actually specify a single evidence accumulation model, and correspondingly, they also did not apply any model to their data. This is problematic, as different evidence accumulation models will make different predictions, depending on the specific model and its parameter values. Moreover, the recognition heuristic on its own does not make predictions about decision times in the memory paradigm (see also Gigerenzer & Goldstein, 2011, for a discussion).

In the memory paradigm, decision times are subject, at least, to the following: the time it takes to read alternatives' names, the time it takes to judge alternatives as recognized or unrecognized, the time it takes to retrieve cues about the alternatives, the time it takes to make a decision as to which alternative to pick, and the time it takes to press a key. In addition a person's intentions (e.g., to respond as quickly as possible) can affect decision times. As a result, decision time predictions warrant not only a model of decision making, but also models of how decision processes interplay with other processes. The recognition heuristic, as formulated by Goldstein and Gigerenzer (2002), remains silent about this interplay; and so do, in fact, most other accessibility-based models of decision making that have been tested in the memory paradigm, including the evidence accumulation and parallel constraint satisfaction models that Hilbig and Pohl (2009) focused on.[1]

Overview

In this chapter, we will model the respective contributions of perceptual, memory, decision, intentional, and motor processes by quantitatively specifying a number of the process hypotheses that have been formulated in the literature in a *cognitive architecture*. A cognitive architecture is a quantitative theory that applies to a broad array of behaviors and tasks, formally integrating theories of memory, perception, action, and other aspects of cognition (for an introduction to cognitive architectures, see, e.g., Gluck, 2010). Among the architectures developed to date (e.g., EPIC, Meyer & Kieras, 1997, and SOAR, A. Newell, 1992), the ACT-R architecture (Anderson et al., 2004) provides perhaps the most detailed account of the various processes that may play a role in accessibility-based decisions. ACT-R has been successfully used to explain phenomena in a variety of fields, ranging from list memory (Anderson et al., 1998), visuospatial working memory (Lyon, Gunzelmann, & Gluck, 2008), diagnostic

[1] *The recognition heuristic has been proposed for the kind of memory-based decisions that are the focus of this chapter (see Figure 5.1; e.g., Gigerenzer & Goldstein, 2011; Goldstein & Gigerenzer, 2002). Using another (i.e., not memory-based) paradigm, Glöckner and Bröder (2011) tested decision time hypotheses they derived from Glöckner and Betsch's (2008) parallel constraint satisfaction model against decision time hypotheses they derived from the recognition heuristic. The testing of these decision time hypotheses represents progress over past studies. However, also these hypotheses fall short of the type of quantitative decision time predictions we advocate. First, on their own, both the recognition heuristic and the parallel constraint satisfaction model remain mute about the interplay of decision, memory, intentional, and motor processes on which decision times in the memory paradigm depend. Second, Glöckner and Bröder's hypotheses concerning decision times are not based on absolute decision times, but on contrast predictions (i.e., one decision strategy will take n times longer than the other).*

reasoning (Mehlhorn et al., 2011; see Chapter 2 of this thesis), and probability learning (Gaissmaier, Schooler, & Rieskamp, 2006), to flying (Gluck, Ball, & Krusmark, 2007), driving (Salvucci, 2006), and the teaching of thousands of children in U.S. high schools with tutoring systems (Ritter, Anderson, Koedinger, & Corbett, 2007). Here, we will use ACT-R to implement 39 process models. These models are the recognition heuristic, as well as various other noncompensatory and compensatory decision strategies, including models that incorporate central aspects of integration, connectionist, evidence accumulation, and race models. In a model competition, we will test the 39 process models' ability to predict people's decisions and decision times in the memory paradigm.

Before we start, three comments are warranted. First, the goal of this chapter is not so much to advocate any particular process model, but rather, using the debate about the recognition heuristic as a case study, to provide a methodological primer on how architectures like ACT-R can be used to lend precision to the theorizing about decision processes. That is, while we also test process models against each other, the model competition's objective is to illustrative methodological principles, and not necessarily to identify the very best model. For those interested in identifying the best model, the main contribution of this chapter is, perhaps, to provide 39 precisely specified process models, cast into the computer code of a detailed cognitive architecture, and ready to be tested in studies beyond the limited data we use here.

Second, there are many research programs that are built around quantitative models (e.g., Busemeyer & Townsend, 1993; Ratcliff & Smith, 2004; Rumelhart, McClelland, & the PDP Research Group, 1986). Certainly, our critique of the lack of specification of process hypotheses only applies to these models to the extent that they remain silent about the interplay of perceptual, memory, decision, intentional, and motor processes. Moreover, we are not the first who discuss decision strategies such as the recognition heuristic and related models in the context of ACT-R or other architectures (Dougherty et al., 2008; Gaissmaier, Schooler, & Mata, 2008; Hertwig et al., 2008; Marewski & Schooler, 2011; Nellen, 2003; Schooler & Hertwig, 2005; Van Maanen & Marewski, 2009).

Third, while it is possible to test evidence accumulation, the recognition heuristic, and other models against each other without implementing these models in a cognitive architecture, such *direct* model comparisons are not without problems, because these models tend to be specified at different levels of description and computational precision, resulting in different levels of detail and precision of the models' predictions. For instance, many evidence accumulation models are specified mathematically and include several free parameters (e.g., Ratcliff & Smith, 2004). The recognition heuristic, in turn, consists of a verbally formulated if-then statement. (If one alternative is recognized, then choose the recognized alternative.) While the parameterized evidence accumulation models can yield predictions about decision time distributions, on its own the recognition heuristic's if-then-statement does not predict such distributions. Much the same can be said with respect to comparisons of other models, including the aforementioned parallel constraint satisfaction and classic integration models. By implementing models of different levels of description and

specificity in *one* architectural modeling framework, we make the models and their predictions comparable, providing a basis for future model tests beyond the ones we will provide below.

The chapter is structured as follows. First, we will describe the experimental data we used to test the models. Second, we will explain the methodological principles guiding our modeling. Third, we will provide an overview of ACT-R as well as of the models we implement. Fourth, we will illustrate how these models' ability to predict people's decisions and decision times can be tested.

Experimental Data

We developed models for memory-based decisions about city size, which is the task most studies on the recognition heuristic have used (Figure 5.1). Specifically, we reanalyze Pachur et al.'s (2008) Experiments 1 and 2.[2] These experiments are well-suited for our purposes, because they entail good control over peoples' recognition and cue-knowledge, this way simplifying our modeling exercise.

Summary of Pachur et al.'s (2008) Pre-studies

To create stimulus materials for their experiments, Pachur et al. (2008) conducted pre-studies wherein they presented participants with names of British cities and had them indicate whether they had heard or seen the names prior to participating in the study, that is, whether they recognized them. Six highly recognized and 10 poorly recognized cities (*R cities* and *U cities*, respectively) were selected as stimuli. Pachur et al. also surveyed what people thought were useful cues for inferring the cities' sizes to establish a stimulus set of cues. These cues were whether a city had significant industry (*industry cue*), an international airport (*airport cue*), or a premier league soccer team (*soccer cue*).

Summary of Pachur et al.'s (2008) Experiment 1

Learning task

The experiment was run with a new group of participants (N = 40, 19 females; mean age = 24.6 years). The experiment started with a *learning task* (as, e.g., used by Bröder & Eichler, 2006; Bröder & Schiffer, 2003), in which participants were taught the three cues about the six R cities. During learning, cities and cues were presented repeatedly in a random order until participants correctly recalled all cities' values on the cues. Table 5.1 summarizes the cues.

[2] *When the article on which this chapter is based was accepted for publication, a part of Pachur et al.'s (2008) data had never been published. This was the case for the reaction times recorded in Pachur et al's experiments, which are modeled using ACT-R below. After the article's acceptance for publication, the authors learned about a new (then still unpublished) manuscript by Pachur (2011), in which an analysis of the reaction times is reported.*

Table 5.1 Cues taught in the learning tasks of Experiments 1 and 2.

Cue	City					
	Aberdeen	Bristol	Nottingham	Sheffield	Brighton	York
Industry	+	+	+	+	+/- [a]	+/- [a]
Airport	+	+	-	-	-	-
Soccer	+	+	+	+	-	-

Note. + = positive cue value. - = negative cue value. [a] The design of Experiment 1 and 2 differed slightly. In Experiment 1, Pachur et al. (2008) taught participants positive values on the industry cue for Brighton and York. In Experiment 2, Pachur et al. taught participants negative values on the industry cue for Brighton and York.

Decision task

After having learned the cues, participants performed the *decision task*. In this task, 120 pairs of British cities were presented on a computer screen (one city on the left side of the screen, the other on the right). Participants were instructed to choose the one with more inhabitants by pressing a key (see Figure 5.1).

For each *trial*, a pair of cities was drawn at random from three types of city pairs. In the main type (i), six R cities that were mostly recognized in the pre-studies were combined with 10 cities that were mostly unrecognized in the pre-studies, yielding 60 *RU pairs*. These 60 pairs were critical for Pachur et al.'s (2008) and our purposes, because they were most likely to allow people to apply the recognition heuristic. We used these pairs to test our models. To balance the presentation frequency of the R and U cities as much as possible, (ii) there were 30 filler pairs consisting of two cities that were mostly unrecognized in the pre-studies (*UU pairs*) as well as (iii) 30 filler pairs consisting of two recognized cities (*RR pairs*).

Recognition task

The decision task was followed by a *recognition task*. Participants were presented all cities in a random order and had to indicate for each city whether they had heard of it before participating in the experiment. The purpose of this recognition task was to make sure that the RU pairs, which were identified based on the pre-studies, also represented RU pairs for the participants of Experiment 1, whose recognition judgments were likely to be similar but not identical to the recognition judgments made in the pre-studies. We used participants' responses in this task to model their recognition of cities.

Cue-memory task

After the recognition task, participants performed a *cue-memory* task in which they had to reproduce the cue values ("yes" or "no") they had learned for the six R cities in the learning task. If they could not recall the correct values, they were allowed to respond

"don't know". The purpose of this task was to test how well participants remembered the cues they were taught. We used participants' responses in this task to model their retrieval of cues; for instance, whether they believed a city to have an airport.

Summary of Pachur et al.'s (2008) Experiment 2

In Experiment 2 (N = 40; 25 females; mean age = 25.2 years), for two cities the positive values on the industry cue were replaced by negative ones, such that recognition was contradicted by three negative cues (see Table 5.1). In all other respects, Experiment 2 was identical to Experiment 1.

Model-Testing Approach: Methodological Principles

To strengthen our modeling efforts, we embraced five methodological principles.

Nested modeling

Any new model should be related to its own precursor (e.g., including it as special cases) and should be tested on data that the old model was able to account for (Grainger & Jacobs, 1996; Jacobs & Grainger, 1994). Our models implement the qualitative hypotheses discussed in the literature in a stepwise, nested fashion, and are tested on Pachur et al.'s (2008) data.

Competitive modeling

A model's ability to account for data should not be evaluated in isolation, but in model comparisons (e.g., Fum, Del Missier, & Stocco, 2007; Gigerenzer & Brighton, 2009; Marewski, Schooler, & Gigerenzer, 2010). In such comparisons, a model's ability to account for data can be compared to that of competing models. For instance, this way it is possible to learn that no model accounts for the data perfectly, but some account for them better than others. This way it is also possible to establish benchmarks in model evaluation; for example, a new model should be able to account for data better than previously existing models that are already known to account well for those data. Unfortunately, this competitive approach to model testing has rarely been taken in recognition heuristic research (but see Glöckner & Bröder, 2011; Marewski et al., 2009; Marewski, Gaissmaier, Schooler et al., 2010; Pachur & Biele, 2007, for exceptions). Here, we test all models competitively.

Constrained modeling

Models should be tested by constraining their parameters in separate tasks (Anderson, 2007; A. Newell, 1990). We calibrated all models' free parameters to the tasks of Experiment 1, using a stepwise procedure to constrain the parameter space. Specifically,

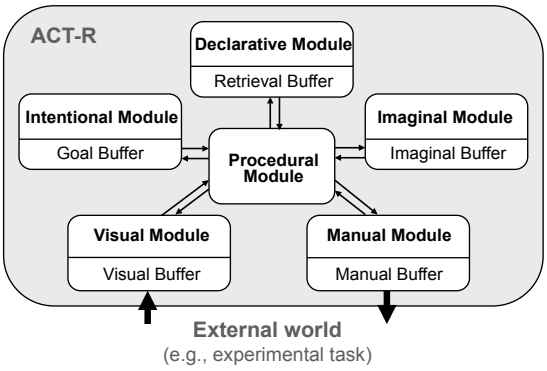


Figure 5.2
The organization of ACT-R. Note that the modules of the architecture have been mapped onto brain regions, enabling detailed process predictions of functional magnetic resonance imaging (fMRI) data (see e.g., Anderson, Fincham, Qin, & Stocco, 2008). While it is beyond the scope of this chapter to test fMRI predictions, we would like to point out that all models reported in this chapter actually allow making such predictions, inviting future model tests.

we *first* fitted the parameters associated with recognition and cue retrieval on data of the recognition and cue-memory tasks of Experiment 1, creating separate ACT-R models of recognition and cue retrieval. With these parameters *fixed*, we then estimated the remaining parameters from participants' decisions and decision times in the decision task of Experiment 1 (see Supplementary Online Material A).

Predictive modeling

We use the term "predicting" (or "generalizing") to refer to situations in which a model's free parameters are fixed such that they cannot adjust to the data on which the model is tested. In contrast, we reserve the term "fitting" (or "calibrating") to refer to situations in which a model's parameters are allowed to adapt to the data. Predicting data well lends credence to a model and is one standard by which models should be evaluated (e.g., Busemeyer & Wang, 2000; Marewski & Olsson, 2009; Pitt et al., 2002; Roberts & Pashler, 2000). We used the parameters fitted on Experiment 1 to predict behavior in Experiment 2.[3]

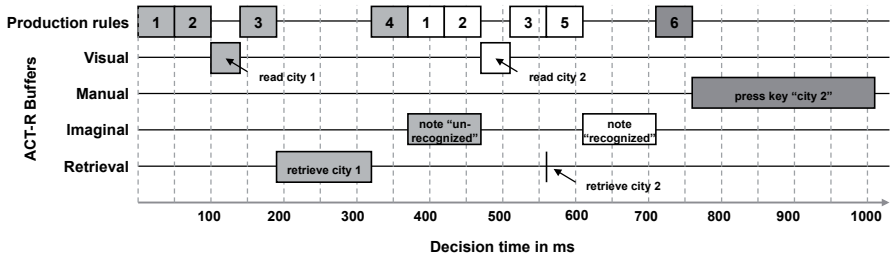
Distributional modeling

Rather than just predicting means of behavioral data, we strive to predict the associated distributions, which further helps evaluating our ACT-R models' ability to account for human data (for a related approach, see Ratcliff & Smith, 2004). Next, we will turn to ACT-R and these models.

Thirty-Nine ACT-R Models of Inference

ACT-R describes human cognition as a set of independent modules that interact through a production system, the procedural module (Figure 5.2). The production system consists of *production rules* (i.e., if–then rules) whose conditions (i.e., the "if"

[3] *The participants of Pachur et al.'s (2008) experiments were recruited and tested in the same laboratories.*



Production rules

1 **If** there is visual information on the screen that has not been attended before
Then move the attention to this information

2 **If** the information on the screen is attended
Then read and encode it in the visual buffer

3 **If** the name of a city has been read
Then attempt to retrieve a chunk representing that city name

4 **If** no chunk representing the city name (here city 1) can be retrieved
Then encode in the imaginal buffer that this city is unrecognized

5 **If** a chunk representing the city name (here city 2) can be retrieved
Then encode in the imaginal buffer that this city is recognized

6 **If** a recognized and an unrecognized city are encoded in the imaginal buffer
And the goal is to decide which city is larger
Then press a key to enter the recognized city as larger than the unrecognized city

Figure 5.3 Processing stream for Model 1, one of our implementations of the recognition heuristic. Light grey boxes depict processing an unrecognized city name; white boxes depict processing a recognized city name. Dark grey boxes depict actions related to the response. Note that predicted decision times represent examples; the model's decision time predictions can vary across different decision trials, for instance, as a function of noisy perceptual and motor processes (Supplementary Online Material A). Production rules are stylized representations of the LISP code productions rules that have been used to implement the models in ACT-R.

parts of the rules) are matched against the modules. If the conditions of a production rule are met, then the production rule can fire. In this case, the action specified by the production rule is carried out.

Each module implements different cognitive processes. The *declarative module* allows information storage in and retrieval from declarative memory, the *intentional module* keeps track of a person's goals, and the *imaginal module* holds information necessary to perform the current task. By this token, the imaginal module is comparable to the focus of attention in working memory (Anderson, 2007; Borst et al., 2010; Oberauer, 2002). A *visual module* for perception and a *manual module* for motor actions (e.g., pressing a key on a computer keyboard) are used to simulate interactions with the world. While the different modules can operate in parallel, information within each module can only be processed in a serial manner (Byrne & Anderson, 2001).

In coordinating the modules, the production rules can act only on information that is available in *buffers*, which can be thought of as processing bottlenecks (Salvucci & Taatgen, 2008), linking the modules' contents to the production rules. For instance, the production rules cannot access all contents of the declarative module, but only the part of information that is currently available in the retrieval buffer.

ACT-R distinguishes a *symbolic* and a *subsymbolic system*. The symbolic system is composed of the production rules as well as the modules and buffers. Access to the information stored in the modules and buffers is determined by the subsymbolic system. This system is cast as a set of equations and determines, for instance, the timing of memory retrieval. Before turning to these equations, let us provide two examples of the ACT-R models we implemented.

Implementing Accessibility-Based Decision Strategies in ACT-R: Two Examples

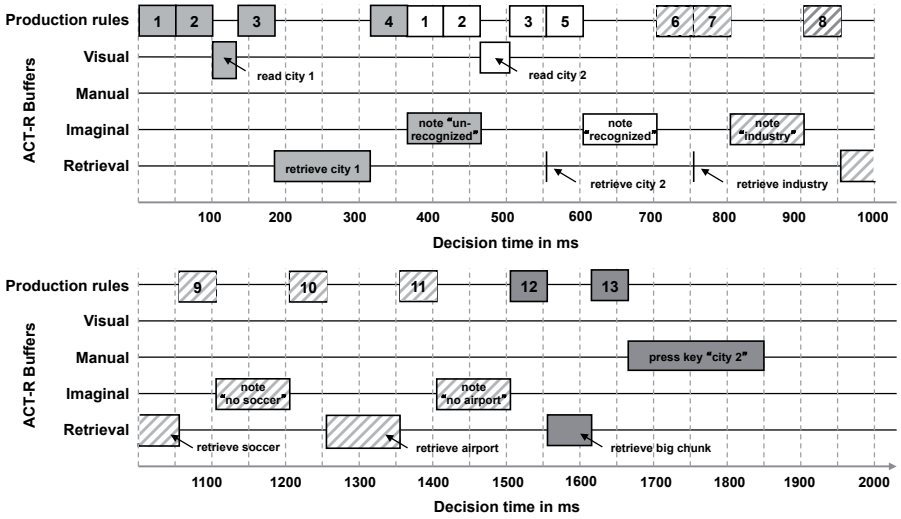
Our ACT-R models perform the same decision task as Pachur et al.'s (2008) experimental participants: They "read" the city names off the computer screen, process them, decide which city is larger, and enter the response by "pressing" a key.

Figure 5.3 shows the processing stream of Model 1, which is one of our implementations of the recognition heuristic. As can be seen, the various processing steps assumed by the model are coordinated by a set of production rules. Specifically, the model assumes that people first read the names of both cities. In doing so, the model attempts to retrieve a memory trace of the cities' names, called a *chunk*. Chunks are facts like "York is a city" or "York has industry" and model people's recognition of city names and their cue knowledge, respectively. If a chunk representing the name of one city can be retrieved, then this city is recognized.[4] In Model 1, retrieving the chunk of one city but not the chunk of the other, is sufficient information to enter the recognized city as the larger city.

To compare, Figure 5.4 shows one of the compensatory strategies we implemented. As can be seen, Model 4.H.PN assumes that, after assessing recognition, a person will retrieve chunks about the recognized city, such as the industry cue. The retrieved cues are stored in the imaginal buffer. As we will explain below, from the imaginal buffer the cues spread a memory signal called *activation* to intuitive knowledge that large cities tend to have airports, premier league soccer teams, and significant industry. In the model, this knowledge is labeled *big chunk*. If the big chunk receives sufficient spreading activation from the retrieved cues, then Model 4.H.PN will recall that the recognized city is a large city and enter this city as response. If the big chunk's activation is too weak, then the big chunk will not be retrieved. Consequently, the model has no reason to assume that the recognized city is large and will respond with the unrecognized city. The assumption is that such subsymbolic processes describe how people make implicit and intuitive, rather than explicit, deliberate judgments.

As can be seen by comparing the x-axes of Figures 5.3 and 5.4, decision times are longer in Model 4.H.PN than in Model 1, because Model 4.H.PN assumes more processing steps than Model 1. In what follows, we give a short overview of the subsymbolic processes that determine the timing of the processing steps in these and all other models.

[4] *In modeling recognition, we follow Anderson et al. (1998) and Schooler and Hertwig (2005) in assuming that a chunk's retrieval implies recognizing it.*



Production rules

1–5 **Identical to in Model 1**

6 **If** a recognized and an unrecognized city are encoded in the imaginal buffer
And it has not yet been probed whether the recognized city has significant industry
Then try to retrieve information indicating whether the city has such industry

7 **If** information that the recognized city has significant industry has been retrieved
Then encode in the imaginal buffer that this city has significant industry

8 **If** a recognized and an unrecognized city are encoded in the imaginal buffer
And it has not yet been probed whether the recognized city has a premier league soccer team
Then try to retrieve information indicating whether it has such a team

9 **If** information is retrieved that the recognized city has no premier league soccer team
Then encode in the imaginal buffer that the recognized city the has no premier league soccer team

10 **If** a recognized and an unrecognized city are encoded in the imaginal buffer
And it has not yet been probed whether the recognized city has an international airport
Then try to retrieve information indicating whether it has such an airport

11 **If** information can be retrieved that the recognized city has no international airport
Then note in the imaginal buffer that the recognized city has no international airport

12 **If** a recognized and an unrecognized city are encoded in the imaginal buffer
And information indicating whether the recognized city has industry, soccer and an airport
Then try to retrieve the big chunk

13 **If** the big chunk can be retrieved
Then press a key to enter the recognized city as larger than the unrecognized city

Figure 5.4 Processing stream for Model 4.H.PN. Light grey boxes depict processing an unrecognized city name; white boxes depict processing a recognized city name. Striped boxes depict actions related to the retrieval of cues. Dark grey boxes depict actions related to the response. Note that predicted decision times represent examples; the model's decision time predictions can vary across different decision trials, for instance, as a function of noisy perceptual and noisy motor processes, or as a function of whether to-be-retrieved cues are positive, negative, or unknown (Supplementary Online Material A). As we explain in detail below, also the order in which cues are processed (i.e., productions 6-11) will vary across trials (see also Footnote 7). Production rules are stylized representations of the LISP code productions rules that have been used to implement the models in ACT-R.

Subsymbolic Memory Processes Assumed by ACT-R

Access to chunks such as "York is a city" or "York has industry" is determined by the chunk's activation (Lovett et al., 2000). The *activation*, A_i, of chunk i (e.g., a city or

a cue) reflects the likelihood that the chunk will be needed in the future (Anderson & Schooler, 1991) and is determined by three components—the chunk's *base-level activation*, B_i, the *spreading activation* the chunk receives from the current context, S_i, and a *noise* component, ε:

$$A_i = B_i + S_i + \varepsilon \tag{5.1}$$

The first component that influences a chunk's activation, A_i, its base-level activation, B_i, reflects the chunk's past usefulness:

$$B_i = \ln\left(\sum_{k=1}^{n} t_k^{-d}\right) \tag{5.2}$$

where n is the number of presentations of chunk i, t_k is the time since the k^{th} presentation, and d is a decay parameter. Consequently, the more often a city name or a cue was encountered (e.g., in an experimental task) and the more recent these encounters were, the higher the city's or cue's activation.[5]

The second component that influences a chunk's activation, A_i, spreading activation, S_i, reflects the chunk's usefulness in the current context. The amount of spreading activation is determined by the chunk's association to other chunks that are currently stored in the buffers (Anderson & Lebiere, 1998). In our models, reading a city name and encoding it in the imaginal buffer would, for example, increase the likelihood of a cue associated with this city being needed. The city would spread activation to the cue as described by Equation 5.3:

$$S_i = \sum_j W_j S_{ji} \tag{5.3}$$

where cue i receives spreading activation, S_i, from city j. The amount of spreading activation S_i is determined by the *associative strength*, S_{ji}, between i and j, which is weighted by the *source activation*, W_j, of j in the imaginal buffer. The associative strength, S_{ji}, between chunks is approximated with

$$S_{ji} = S - \ln(fan_{ji}) \tag{5.4}$$

where S is a parameter for the *maximum associative strength* between chunks and fan_{ji} is the number of chunks i that are associated with a chunk j. Consequently, the more cues are associated with a city in memory, the lower the associative strength between the city and each of the cues.

[5] *In modeling Pachur et al.'s (2008) experimental tasks, we assume the base level activations (i.e., of the cities, cues, and the big chunk) to vary only across the time it takes to make a decision in a trial in the decision task, as well as across the times it takes to make a judgment in a trial of the recognition and cue memory tasks, respectively. For instance, decisions that take a long time are more likely to allow for the base level activations to decay away than decisions that are made quickly. For simplicity, we re-set the base level activations to their initial values (see Supplementary Online Material A) each time a new trial was presented. For example, upon presentation of a trial consisting of the cities of York and Stockport, the base level activations would be allowed to vary until a decision is made for that trial. For the next trial, say the cities of Bristol and Poole, the base level activations would first be re-set to their initial values, and then be allowed to vary until a decision is made in that trial.*

The third component that influences a chunk's activation, A_i, is the retrieval noise, ε. It is added to the activation of a chunk when a retrieval request is made. With s being a free parameter, ε is generated from a logistic distribution with a mean of zero and a variance of

$$\sigma^2 = \frac{\pi^2}{3} s^2 \tag{5.5}$$

Only chunks that exceed a certain amount of activation, A_i, as defined by the *retrieval threshold*, τ, can be retrieved. For instance, only cues with activations falling above τ would be retrieved. The retrieval probability, p, is:

$$p = \frac{1}{1 + e^{\frac{\tau - A_i}{s}}} \tag{5.6}$$

If a chunk i can be retrieved, the time required for the retrieval is determined by the *latency factor*, F, and the activation of the chunk, A_i:

$$Retrieval\ \ Time = Fe^{-A_i} \tag{5.7}$$

Thus, the more strongly city names and cues are activated in memory, the faster they can be retrieved.

If no chunk matches a retrieval request or if the matching chunk with the highest activation is below the retrieval threshold, a retrieval failure will occur. For example, reading the name of an unknown city will result in a retrieval failure. The time it takes to note such a failure is:

$$Retrieval\ \ Failure\ \ Time = Fe^{-\tau} \tag{5.8}$$

Detailed Description of the 39 Models

The above-described subsymbolic memory processes as well as the corresponding parameter values are identical in all models and the models also do not differ with respect to the perceptual and motor processes they assume (Supplementary Online Material A).

However, the models *do* differ with respect to the decision processes. In implementing these processes, we had to make a series of assumptions, for instance, about the order in which people will assess recognition as opposed to cues. All assumptions are grounded in the decision, memory, and ACT-R literatures. Often, however, these literatures offer more than one plausible assumption. Following the principle of competitive modeling, we dealt with such competing assumptions by creating different models to implement them, which allowed us to test the assumptions against each other. Following the principle of nested modeling, we additionally combined part of these assumptions with each other, resulting in 39 models. These models are summarized in Tables 5.2 and 5.3.

Table 5.2 Overview of the perception and memory processes used in the 39 models.

	Retrieve and encode city names	Retrieve positive cues	Retrieve negative cues	Number of retrieved cues [a]	Retrieved cues can be forgotten	Encode cues in the imaginal buffer
Model 1 class: *Stopping and decision rules noncompensatory—simple model*						
Model 1	X			0		
Model 2 class: *Stopping rule compensatory, decision rule noncompensatory—simple models*						
Model 2.PN	X	X	X	3		
Model 2.P	X	X		3		
Model 3 class: *Stopping rule compensatory, decision rule noncompensatory—simple models*						
Model 3.PN	X	X	X	3		X
Model 3.P	X	X		3		X
Model 1&3 class: *Stopping rule noncompensatory and compensatory, decision rule noncompensatory—race models*						
Model 1&3.PN	X	X	X	0 to 3		X
Model 1&3.P	X	X		0 to 3		X
Model 1&3.PN.F	X	X	X	0 to z [b]	X	X
Model 1&3.P.F	X	X		0 to z [b]	X	X
Model 4 class: *Stopping rule compensatory, decision rule compensatory—simple models*						
Model 4.H.PN	X	X	X	3		X
Model 4.H.P	X	X		3		X
Model 4.L.PN	X	X	X	3		X
Model 4.L.P	X	X		3		X
Model 1&4 class: *Stopping rule noncompensatory and compensatory, decision rule noncompensatory and compensatory—race models*						
Model 1&4.H.PN	X	X	X	0 to 3		X
Model 1&4.H.P	X	X		0 to 3		X
Model 1&4.H.PN.F	X	X	X	0 to z [b]	X	X
Model 1&4.H.P.F	X	X		0 to z [b]	X	X
Model 1&4.L.PN	X	X	X	0 to 3		X
Model 1&4.L.P	X	X		0 to 3		X
Model 1&4.L.PN.F	X	X	X	0 to z [b]	X	X
Model 1&4.L.P.F	X	X		0 to z [b]	X	X
Model 5 class: *Stopping rule compensatory, decision rule compensatory—simple models*						
Model 5.1.PN	X	X	X	1 to 3		X
Model 5.1.P	X	X		1 to 3		X
Model 5.2.PN	X	X	X	2 to 3		X
Model 5.2.P	X	X		2 to 3		X
Model 5.3.PN	X	X	X	3		X
Model 5.3.P	X	X		3		X
Model 1&5 class: *Stopping rule noncompensatory and compensatory, decision rule noncompensatory and compensatory—race models*						
Model 1&5.1.PN	X	X	X	0 to 3		X
Model 1&5.1.P	X	X		0 to 3		X
Model 1&5.1.PN.F	X	X	X	0 to z [b]	X	X
Model 1&5.1.P.F	X	X		0 to z [b]	X	X
Model 1&5.2.PN	X	X	X	0 to 3		X
Model 1&5.2.P	X	X		0 to 3		X
Model 1&5.2.PN.F	X	X	X	0 to z [b]	X	X
Model 1&5.2.P.F	X	X		0 to z [b]	X	X
Model 1&5.3.PN	X	X	X	0 to 3		X
Model 1&5.3.P	X	X		0 to 3		X
Model 1&5.3.PN.F	X	X	X	0 to z [b]	X	X
Model 1&5.3.P.F	X	X		0 to z [b]	X	X

Note. PN = Positive and negative cues. P = positive cues. F = forgetting cues. [a] As retrieved cues, we count all (positive, negative, and unknown) cue values that have been probed in memory. [b] The maximum number of retrieved cues is variable, because cues can be retrieved again when they are forgotten. For a description of the parameter settings, see Supplementary Online Material A; for model codes see http//www.ai.rug.nl/~katja/models.

As can be seen in Table 5.2, the labeling of the models is organized around eight main classes: the Model 1, 2, 3, 4, 5, 1&3, 1&4, and 1&5 class, with each class embodying different sets of assumptions. Specifically, as we will discuss in more detail below, the Model 1 class implements what one may loosely term noncompensatory processes; the Model 2 and 3 classes implement noncompensatory and compensatory processes; and the Model 4 and 5 classes implement only compensatory processes.[6] The Model 1&3, 1&4, and 1&5 classes were generated by partially combining the Model 1, 3, 4, and 5 classes with each other. For example, combining Model 1 and Model 3 resulted in the Model 1&3 class. In what follows we will describe the models in more detail. Complete model codes can be downloaded from http://www.ai.rug.nl/~katja/models.

Primacy of recognition

As a first processing step, all models read the city names (in Table 5.2, column labeled retrieve and encode city names). If they can retrieve a city, they encode it as recognized in the imaginal buffer. If they cannot retrieve a city, they encode it as unrecognized. Put differently, we assume that people will first assess their recognition of the city names before retrieving further cues. This assumption is grounded in our experimental setup, in which participants were shown the city names but no cues (Figure 5.1). Moreover, this assumption is consistent with the literature, which suggest that familiarity (i.e., recognition) arrives on the mental stage earlier than recollection (e.g., Gronlund & Ratcliff, 1989; Hertwig et al., 2008; Hintzman & Curran, 1994; McElree, Dolan, & Jacoby, 1999; Pachur & Hertwig, 2006; Volz et al., 2006).

The models differ in the steps that are executed after recognition has been assessed. Whereas Model 1 bases decisions only on recognition, the remaining 38 models additionally retrieve cues. In all of these 38 models, the retrieval of cues is instantiated by three sets of production rules, which attempt to retrieve a city's value on the soccer, industry, and airport cues, respectively. If such a retrieval attempt is successful, the cue value is retrieved from memory. If the attempt is not successful (a retrieval failure occurs), the value of this cue is unknown to the model. (For simplicity, in both cases we speak of the respective cues as having been "retrieved", because, even if the cue value is unknown, the cue has been probed in memory.) Which production fires first, and correspondingly, which cue is retrieved first, is determined at random. We implemented this random cue retrieval order, because during the learning task all cues were presented equally often in random order until they were remembered perfectly, making it equally likely for a person to remember that a city has a premier league soccer team, a significant industry, or an international airport, respectively.

[6] *Note that we use the terms "noncompensatory" and "compensatory" in a loose sense to help readers to map the verbal descriptions of our ACT-R models to the existing literature on the recognition heuristic. However, there is, perhaps, no one-to-one mapping. A more adequate way of thinking about our models might be that they represent the dimension recognition-based versus cue-based, which in fact also reflects the dichotomy on which the controversy about noncompensatory versus compensatory process models of decision making has focused on in the recognition literature. We would like to point interested readers to our model codes for precise information on what our models look like.*

Table 5.3 Overview of the decision process and its outcome for the 39 models.

	Information used in the decision process				Outcome of the decision process		
	Use recognition to choose between cities	Use cues to choose between cities	Use cues via sub-symbolic system	Use cues via symbolic system	Always choose recognized city	Sometimes choose un-recognized city	Decision time is influenced by cues
Model 1 class: *Stopping and decision rules noncompensatory—simple model*							
Model 1	X				X		
Model 2 class: *Stopping rule compensatory, decision rule noncompensatory—simple models*							
Model 2.PN	X				X		X
Model 2.P	X				X		X
Model 3 class: *Stopping rule compensatory, decision rule noncompensatory—simple models*							
Model 3.PN	X				X		X
Model 3.P	X				X		X
Model 1&3 class: *Stopping rule noncompensatory and compensatory, decision rule noncompensatory—race models*							
Model 1&3.PN	X				X		X
Model 1&3.P	X				X		X
Model 1&3.PN.F	X				X		X
Model 1&3.P.F	X				X		X
Model 4 class: *Stopping rule compensatory, decision rule compensatory—simple models*							
Model 4.H.PN		X	X			X	X
Model 4.H.P		X	X			X	X
Model 4.L.PN		X	X			X	X
Model 4.L.P		X	X			X	X
Model 1&4 class: *Stopping rule noncompensatory and compensatory, decision rule noncompensatory and compensatory—race models*							
Model 1&4.H.PN	X	X	X			X	X
Model 1&4.H.P	X	X	X			X	X
Model 1&4.H.PN.F	X	X	X			X	X
Model 1&4.H.P.F	X	X	X			X	X
Model 1&4.L.PN	X	X	X			X	X
Model 1&4.L.P	X	X	X			X	X
Model 1&4.L.PN.F	X	X	X			X	X
Model 1&4.L.P.F	X	X	X			X	X
Model 5 class: *Stopping rule compensatory, decision rule compensatory—simple models*							
Model 5.1.PN	X[a]	X		X		X	X
Model 5.1.P	X[a]	X		X	X		X
Model 5.2.PN	X[a]	X		X		X	X
Model 5.2.P	X[a]	X		X	X		X
Model 5.3.PN	X[a]	X		X	X[b]	X[b]	X
Model 5.3.P	X[a]	X		X	X		X
Model 1&5 class: *Stopping rule noncompensatory and compensatory, decision rule noncompensatory and compensatory—race models*							
Model 1&5.1.PN	X	X		X		X	X
Model 1&5.1.P	X	X		X	X		X
Model 1&5.1.PN.F	X	X		X		X	X
Model 1&5.1.P.F	X	X		X	X		X
Model 1&5.2.PN	X	X		X		X	X
Model 1&5.2.P	X	X		X	X		X
Model 1&5.2.PN.F	X	X		X		X	X
Model 1&5.2.P.F	X	X		X	X		X
Model 1&5.3.PN	X	X		X	X[b]	X[b]	X
Model 1&5.3.P	X	X		X	X		X
Model 1&5.3.PN.F	X	X		X	X[b]	X[b]	X
Model 1&5.3.P.F	X	X		X	X		X

Note. PN = Positive and negative cues. P = positive cues. F = forgetting cues. [a] Models of the Model 5 class use recognition to decide between cities if they cannot reach their decision criterion of *D* cues. [b] In Experiment 1, the PN versions of the Model 5.3 and 1&5.3 classes always choose recognized cities, because these models require at least three negative cues to choose unrecognized cities (*D* = 3). In Experiment 2, the models sometimes choose unrecognized cities, because in this experiment cases with three negative cues occurred (see Table 5.1).

Positive and negative cues

It has been argued that people are more likely to use positive cues rather than negative ones (Dougherty et al., 2008; Glöckner & Bröder, 2011). We incorporated this hypothesis in the models. As can be seen in Table 5.2, except for Model 1, which does not retrieve any cues, for all models we created two versions, one that retrieves positive and negative cues (labeled *PN version*, e.g., Model 2.PN) and one that retrieves only positive cues (labeled *P version*; e.g., Model 2.P). Note that retrieving negative cues is not necessary to decide in favor of unrecognized cities (see descriptions of Model 4 and Model 1&4 below). Also note that we assume positive cues to be more strongly activated and therefore to be retrieved faster than negative ones (Supplementary Online Material A).

Model 1, 2, and 3 classes: Models with noncompensatory decision rules

As mentioned above, Model 1 assesses recognition only, always inferring recognized cities to be larger than unrecognized ones (see Table 5.3). Also Models 2.PN, 2.P, 3.PN, and 3.P always infer recognized cities to be larger than unrecognized ones. Yet, these four models additionally retrieve cues. Adding yet another processing step, Models 3.PN and 3.P do not only retrieve the cues, but also encode their values (e.g., in Model 3.PN: positive, negative, or unknown) in the imaginal buffer. This encoding is time costly (see Supplementary Online Material A, imaginal-delay), but it allows the cues to be available in working memory (i.e., in the imaginal buffer) for further processing steps and to spread activation to other information in memory.

In the terminology often used to describe the recognition heuristic and related heuristics, in Models 2.PN, 2.P, 3.PN, and 3.P what one may term "compensatory processes" govern the models' *stopping rules*, that is, the models' rules for deciding when to stop information retrieval. What one my term "noncompensatory processes" direct the models' *decision rules*, that is, the rules on how available information is used to make a decision. In Model 1, in contrast, both the stopping and the decision rules are noncompensatory.

Model 1 corresponds to what we deem to be the simplest recognition heuristic implementation; Models 2.PN, 2.P, 3.PN, and 3.P in turn, also implement the recognition heuristic, but incorporate more recent hypotheses about the heuristic's stopping rule (Gigerenzer & Goldstein, 2011; Pachur et al., 2008). For example, the compensatory stopping rule in Model 3.PN will cause the model to stop information retrieval when it has retrieved and encoded the information of all three cues. The noncompensatory decision rule will then cause the model to ignore the cues and to decide based on the recognition of the cities.

Model 4 and 5 classes: Models with compensatory decision rules

The Model 4 and 5 classes implement both compensatory stopping and compensatory decision rules. As such, these models are representatives of the type of decision

strategies that is often discussed as antipode to both the recognition heuristic and related noncompensatory heuristics (e.g., Bergert & Nosofsky, 2007; Bröder & Eichler, 2006; Bröder & Gaissmaier, 2007; Bröder & Schiffer, 2003; Glöckner & Hodges, 2011; Hilbig & Pohl, 2009; Mata et al., 2007; B. R. Newell & Fernandez, 2006; B. R. Newell & Shanks, 2004; Oeusoonthornwattana & Shanks, 2010; Pohl, 2006; Richter & Späth, 2006; Rieskamp & Hoffrage, 2008). Specifically, models of the 4 and 5 classes retrieve the city names and cues and encode them in the imaginal buffer, just as Models 3.PN and 3.P do. However, in contrast to Models 3.PN and 3.P, the Model 4 and 5 classes actually use the cues in the decision rules. We distinguish between two pathways of cue usage: subsymbolic, capturing how people make implicit, intuitive decisions, and symbolic, modeling explicit, deliberate decisions.

Subsymbolic use of cues. In the Model 4 class, the retrieved and encoded cues influence the decision through subsymbolic channels, that is, through spreading activation (Equation 5.3). If for a given city, positive cues are encoded in the imaginal buffer, then these positive cues can spread activation to a chunk, labeled big chunk (Figure 5.4). If the activation is strong enough for the big chunk to cross the retrieval threshold, the big chunk will be retrieved and the model will judge the recognized city as large. If the big chunk does not receive sufficient spreading activation to cross the retrieval threshold, the model chooses the unrecognized city. As explained above, we assume this big chunk to reflect intuitive knowledge implicating that a city is large.

How easily the big chunk will be retrieved varies between the models. In Models 4.H.PN and 4.H.P, the big chunk's base-level activation is *higher* (hence *H*) than the retrieval threshold (Supplementary Online Material A), such that the big chunk is likely to be retrieved. As a result these two models often (but not always) judge recognized cities to be larger than unrecognized ones. In Models 4.L.PN and 4.L.P the big chunk's base-level activation is *lower* (hence *L*) than the retrieval threshold. Therefore, the retrieval of the big chunk will more strongly depend on how much activation is spread from positive cues to the big chunk. Importantly, all variants of Model 4 can decide in favor of unrecognized cities even if no negative cues are available, because such decisions depend on the big chunk, which only receives spreading activation from positive cues.

By assuming subsymbolic spreading activation and intuitive knowledge to be responsible for compensatory decision processes, the Model 4 class implements a central feature of connectionist parallel constraint satisfaction models (Glöckner & Betsch, 2008; Thagard, 1989a, 2000), which Glöckner and Bröder (2011) and others (e.g., Hilbig & Pohl, 2009; Hochman et al., 2010) have argued account for behavior better than the recognition heuristic.

Symbolic use of cues. In the Model 5 class, retrieved and encoded cues influence the decision through symbolic pathways, reflecting more deliberate, explicit decision processes. Specifically, production rules check whether a required number of cues has been retrieved to decide whether the recognized city is larger than the unrecognized one or vice versa. As soon as *D* positive cues have been encoded, the models decide

for the recognized city; as soon as D negative cues have been encoded they decide for the unrecognized city, with D representing the decision criterion. If the models cannot retrieve D cues, they use recognition as their best guess, deciding in favor of the recognized city. This also reflects the hypothesis that it is easier to go with than against recognition when making decisions (Pachur & Hertwig, 2006; Volz et al., 2006). Models 5.3.PN and 5.3.P employ a decision criterion of $D = 3$. The decision criterion of Models 5.2.PN and 5.2.N is $D = 2$. Models 5.1.PN and 5.1.P have the lowest decision criterion, with $D = 1$.

For example, assume Model 5.1.PN infers whether York or Stockport is larger. After judging York as recognized and Stockport as unrecognized, the model retrieves cues. The first retrieved cue has a positive value. Thus, the model decides that York is the larger city. If the first retrieved cue had had a negative value, then the model would have decided that the unrecognized city, Stockport, is larger. If the value of the first cue had been unknown (i.e., attempting to retrieve one cue would have resulted in a retrieval failure), then the model would have continued to retrieve cues, until the decision criterion of $D = 1$ positive or negative cues would have been reached. If all cue values had turned out to be unknown, then the model would have used recognition and decided for York.[7]

In sampling as many cues as needed to reach a decision criterion, the Model 5 class implements a feature of sequential sampling and evidence accumulation models that some have suggested describe behavior better than the recognition heuristic and related noncompensatory heuristics (e.g., Hilbig & Pohl, 2009; Lee & Cummins, 2004; B. R. Newell, 2005; B. R. Newell & Lee, in press). By specifying a decision criterion to decide in favor of unrecognized cities, the Model 5 class also resembles the type of compensatory strategies discussed by Marewski, Gaissmaier, Schooler et al. (2010); which, however, assume no sequential sampling of cues. Finally, by placing equal importance on sampled (i.e., retrieved) cues, the Model 5 class implements a feature of classic unit-weight linear integration strategies (e.g., Dawes, 1979; Dawes & Corrigan, 1974; Einhorn & Hogarth, 1975; Gigerenzer & Goldstein, 1996); but also these classics assume no sequential cue sampling.

[7] *To clarify, the order of cue retrieval has* no *impact on the decisions or decision times in models that retrieve all cues before a decision is made (in the experiments we modeled, these are the Model 2, 3, 4, 5.3 classes). The order of cue retrieval* does *have an impact on the decision and decision times in the Model 5.1, 1&5.1, 5.2, and 1&5.2 classes, because these models require fewer than three cues to be retrieved before a decision is made (*$D = 1$ *and* $D = 2$*, respectively). In these models, the same comparison of cities can lead to different decisions and decision times, depending on cue order. Note that decision times in these models* also *depend on cue order because positive cues will be retrieved faster than negative ones (Supplementary Online Material A), resulting in shorter decision times when positive cues are retrieved than when negative ones are retrieved before a decision is made. Due to the different retrieval times for positive and negative cues, the order of cue retrieval can also impact decision times in the Model 1&3, 1&4, and 1&5.3 classes, even though in these models the decisions do not depend on cue order.*

Model 1&3, 1&4, and 1&5 classes: Race models

We refer to all models described above as *simple models* and distinguish them from *race models* (Logan, 1988).[8] Simple models implement only *one* type of decision process. Race models, in contrast, implement a race between *competing* processes. The outcome of this race determines which process will ultimately be responsible for the decision. Specifically, the Model 1&3 class implements a race between Model 1, that is, the simple noncompensatory process to respond with the recognized city, and Model 3, that is, the compensatory process to retrieve and encode cues. The Model 1&4 class implements a race between the noncompensatory process of Model 1 and the subsymbolic compensatory processes to retrieve, encode, and use cues as assumed by Model 4.[9] The Model 1&5 class implements a race between the noncompensatory process of Model 1 and the symbolic compensatory processes to retrieve, encode, and use cues as assumed by Model 5.

To give an example from the Model 1&3 class, Model 1&3.PN first reads and encodes the city names. After these first steps, a race between responding directly with the name of the recognized city (i.e., as in Model 1) and retrieving and encoding one of the three cues (i.e., as in Model 3.PN) takes place. If a retrieve-cue process wins, the retrieved cue is encoded in the imaginal buffer and the race starts again. This race is repeated either (a) until the model responds with the recognized city before all three cues are retrieved (as in Model 1), or (b) until all three cues are encoded and a decision is made in favor of the recognized city (as in Model 3.PN).

As is explained in detail in Supplementary Online Material B, in all race models, we assume that the respond-with-recognized-city process (i.e., Model 1) competes with all other processes of the respective simple model version (i.e., Model 3, Model 4, or Model 5). Consequently, the more steps are required prior to a decision being made, the more often the respond-with-recognized-city process will compete against other processes. To illustrate this, in the Model 1&4 class, the respond-with-recognized-city process competes not only with the retrieve-cue process, but, once all cues are retrieved, also with the process of retrieving a big chunk (as in the Model 4 class).

---

[8] *In the literature, the terms "race" or "race model" are sometimes used in similar ways as the terms "evidence accumulation" or "sequential sampling models". For instance, Gold and Shadlen (2007) define race models as models where "evidence supporting the various alternatives is accumulated independently to fixed thresholds" (p. 541) and as soon as one of the alternatives reaches the threshold, it is chosen. Applying the race to production rules, we implemented a simplified version of that mechanism, where competing production rules have equal* utilities *(Anderson et al., 2004) and are therefore chosen at random. Put in Golden and Shadlen's terms, the production rules have equal chances of reaching the threshold. We choose this implementation, because we did not want to add additional assumption about the relative speed of the various processes involved. Note that the utilities of the production rules did not change over the experiment (i.e., put in ACT-R's terminology, there was no utility learning). We decided for this implementation because participants (and thus also the models) did not receive feedback during the decision phase of the experiments.*

[9] *Note that in all representatives of the Model 4 and 1&4 classes, cue knowledge will be used for the decision only once all cues have been retrieved from memory. We decided for this implementation, because constraint satisfaction models are usually concerned with the integration of information at one certain point in time (see Mehlhorn & Jahn, 2009, and Wang et al., 2006b, for attempts to extent constraint satisfaction models to sequential reasoning). By letting the models do the implicit evaluation of the alternatives only after all cues have been retrieved, we try to stay as close as possible to constraint satisfaction models as proposed in the decision making literature (e.g., Glöckner & Betsch, 2008).*

Whereas in the Model 1&3 and 1&4 classes potentially all three cues can be retrieved (i.e., if the respond-with-recognized-city process does not win the race prior to retrieving all three cues), in the Model 1&5 class the number of cues that can be retrieved depends on the decision criterion *D*. For example, in Model 1&5.1.PN, which has a decision criterion of $D = 1$ positive or negative cue, the respond-with-recognized-city process competes with the retrieve-cue process until one positive or one negative cue has been retrieved. In Model 1&5.2.PN ($D = 2$) the race continues until two positive or negative cues have been retrieved. In Model 1&5.3.PN ($D = 3$) the race continues until three positive or negative cues have been retrieved. If a model of the Model 1&5 class has retrieved all cues without reaching its decision criterion *D*, it will use recognition as its best guess (as in the Model 5 class).

For all race models, we additionally implemented variants that not only assume a race between noncompensatory recognition and compensatory cue retrieval and usage, but additionally assume that retrieved cues will at times be forgotten, such that these cues have to be retrieved again. These models are marked with an *F* in their name (e.g., Model 1&3.PN.F). The intuition is that the various retrieval, encoding, and decision processes can detract from previously retrieved cues (see Lewandowsky et al., 2009, for a discussion of interference-based forgetting in working memory). Specifically, these models start with a race between responding with the recognized city and retrieving and encoding more cues. As soon as at least two cues have been encoded in the imaginal buffer, an additional race against a forgetting process takes place.[10] If this forgetting process wins the race, the retrieved cues are forgotten (i.e., they are removed from the imaginal buffer). If cues are forgotten, then the race between responding with the recognized city and retrieving and encoding cues takes place again. These processes continue until a decision is made.

As can be seen in Table 5.2, the 1&4 and 1&5 race Model classes consist of 8 and 12 different models, respectively. The large number of models within these race model classes is a product of our principle of nested modeling: Recall that the Model 4 class exists in two versions, L and H, representing low and high activation levels of the big chunk. Likewise, the Model 5 class exists in 3 versions, with each one making different assumptions about the number of cues that will be processed (i.e., $D = 1$, 2, or 3). To spare the reader from having to parse long lists of model names, below we subsume the models from these different versions of the Model 1&4 and 1&5 classes under the labels Model 1&4.L and 1&4.H classes, as well as Model 1&5.1, 1&5.2, and 1&5.3 classes, respectively.

[10] *For simplicity, we implemented the forgetting process by means of production rules. We determined the threshold of two cues based on ad-hoc considerations about the positive skew in the human decision time distribution. The possibility of forgetting cues as soon as two cues have been retrieved and encoded results in an increased upper spread (i.e., visible in the 3rd quartile) of the models' decision time distributions.*

Description of the Data Analyses

Individual Differences

It has been pointed out that people may differ in the strategies they use when making decisions from the accessibility of memories (e.g., Bergert & Nosofsky, 2007; Bröder & Gaissmaier, 2007; Cokely, Parpart, & Schooler, 2009; Gigerenzer & Brighton, 2009; Hilbig, 2008; Marewski et al., 2009; Marewski, Gaissmaier, Schooler et al., 2010; B. R. Newell & Shanks, 2004). For instance, Pachur et al. (2009) provided evidence that processing speed influences people's reliance on recognition.

Also Pachur et al. (2008) interpreted their data as being suggestive of individual differences: While some of their participants always chose recognized cities irrespective of the cues they had been taught, other participants' decisions seemed to have been influenced by these cues (see also Pachur, 2011). In reanalyzing Pachur et al.'s data, we took possible individual differences into account by examining the data separately for (a) those participants who always inferred recognized cities to be larger than unrecognized ones (*recognition group*; $n_{Experiment\ 1} = 25$, $n_{Experiment\ 2} = 19$), and (b) those participants who sometimes inferred unrecognized cities to be larger (*cue group*; $n_{Experiment\ 1} = 15$, $n_{Experiment\ 2} = 21$).

Moreover, we tailored the 39 models to each individual participant in two steps. First, each participant's responses in the recognition and cue-memory tasks were used to model the contents of that participant's declarative memory. That is, we did not give the models perfect knowledge of the cities and cue profiles as shown in Table 5.1 but rather let the models operate on each individual participant's recognition and knowledge, as assessed by the recognition and cue-memory tasks, respectively (see http://www.ai.rug.nl/~katja/models for each participants' knowledge as used by the models). Second, using participants' individual recognition and cue knowledge, all models were run on each participant's trials in the decision task.

Assessing the Correspondence Between the Models' Predictions and the Human Data

For simplicity and following the principle of nested modeling, we assessed the correspondence between the models' predictions and the human data by analyzing these data in the same way Pachur et al. (2008) analyzed the human data. Specifically, we collapsed the human data across participants, calculating means and standard errors for proportions (for decisions) as well as medians and the 1st and 3rd quartiles (for decision times) separately for each of 2x3 categories of comparisons of cities. In Experiment 1, these categories were: the recognized city is associated with (a) one positive cue, (b) two positive cues, or (c) three positive cues, and the recognized city is associated with (a) two negative cues, (b) one negative cue, or (c) zero negative cues. In Experiment 2, the 2x3 categories were: the recognized city is associated with (a) zero, (b) two, or (c) three positive cues and with (a) three, (b) one, or (c) zero negative cues.

In both experiments, the definition of the 2x3 categories was based on the cue profiles participants had been taught in the learning tasks (Table 5.1).[11]

Decisions and decision times produced by the models could vary between individual runs, due to noise and, where applicable, due to the race between different processes. Therefore, to compute the models' predictions, for each participant of Experiments 1 and 2, each model was run 40 times. For each of these 40 runs, we calculated means and standard errors as well as medians and 1st and 3rd quartiles, separately for each of the 2x3 categories of each experiment in an analogous way as for the human data. For each category, the means, standard errors, medians, and quartiles were then averaged across the 40 simulation runs.

Results of the Model-Fitting Competition in Experiment 1

Due to the large number of models, in what follows we will mainly discuss the best models' fits. All models' fits are summarized in Table 5.4 and discussed in more detail in Supplementary Online Material C. Supplementary Online Material C also includes a complete set of graphs of all models' fits.

Recognition group

Figure 5.5 shows the human decisions and decision times in the recognition group as well as the decisions and decision times produced by the Model 1&3 class. Within this model class, Model 1&3.P.F produced the smallest RMSDs (root mean square deviations) to the human data. As can be seen, neither the human decisions nor the model's decisions vary as a function of the cues. At the same time, the human and the model's decision times increase with the number of negative cues, decrease with the number of positive cues and show overall a large spread. Also the three remaining models of the 1&3 class, Models 1&3.PN, 1&3.PN.F, and 1&3.P, fit the decisions and decision times well. These three models are identical to Model 1&3.P.F except that they make no assumptions about the forgetting of cues (Models 1&3.PN, 1&3.P) and/or assume negative cues to be represented in memory (Models 1&3.PN, 1&3.PN.F).

As can be seen in Table 5.4 as well as by comparing Figures 5.5 and 5.6, those representatives of the Model 1&5 class that assume a decision criterion of 3 cues (Model 1&5.3.PN, Model 1&5.3.PN.F, Model 1&5.3.P, Model 1&5.3.P.F) fit the decisions and decision times about as well as the Model 1&3 class. For example, the best-fitting model from the Model 1&5.3 class, Model 1&5.3.P.F, produces basically the same decision time pattern as the best-fitting model from the Model 1&3 class, Model 1&3.P.F, and virtually the same RMSDs. Also those representatives of the

[11] *Note that categories defined by positive cues are not necessarily identical to categories defined by negative cues, because both participants and models may sometimes fail to recall whether a cue is positive or negative (i.e., reflected by unknown cue values in the cue-memory task). For instance, the category "two positive cues" does not necessarily correspond to the category "one negative cue". Yet, most of the time the categories as defined by positive and negative cues are identical, because unknown cue values were very rare in the data (see Pachur et al., 2008). Therefore, the results tend to be similar when plotting the data either as a function of positive cues or as a function of negative cues.*

Table 5.4 Root mean square deviations between the model and the human data in Experiment 1.

	Recognition group		Cue group	
	Decisions (%)	Decision times (ms)	Decisions (%)	Decision times (ms)
Model 1 class: *Stopping and decision rules noncompensatory—simple model*				
Model 1	0	409	9.4 [b]	511
Model 2 class: *Stopping rule compensatory, decision rule noncompensatory—simple models*				
Model 2.PN	0	258	9.4 [b]	355
Model 2.P	0	283	9.4 [b]	376
Model 3 class: *Stopping rule compensatory, decision rule noncompensatory—simple models*				
Model 3.PN	0	357	9.4 [b]	449
Model 3.P	0	379	9.4 [b]	469
Model 1&3 class: *Stopping rule noncompensatory and compensatory, decision rule noncompensatory—race models*				
Model 1&3.PN	0	110	9.4 [b]	219
Model 1&3.P	0	97	9.4 [b]	201
Model 1&3.PN.F	0	73	9.4 [b]	185
Model 1&3.P.F	0	67	9.4 [b]	169
Model 4 class: *Stopping rule compensatory, decision rule compensatory—simple models*				
Model 4.H.PN	10.7 [a]	427	0.9	499
Model 4.H.P	10.7 [a]	477	1.5	518
Model 4.L.PN	58.6 [a]	461	49.4	514
Model 4.L.P	59.1 [a]	511	49.6	534
Model 1&4 class: *Stopping rule noncompensatory and compensatory, decision rule noncompensatory and compensatory—race models*				
Model 1&4.H.PN	1.3 [a]	105	8.1	223
Model 1&4.H.P	1.3 [a]	100	8.1	198
Model 1&4.H.PN.F	.8 [a]	71	8.5	177
Model 1&4.H.P.F	.7 [a]	55	8.6	157
Model 1&4.L.PN	7.3 [a]	109	1.9	218
Model 1&4.L.P	7.3 [a]	95	2.1	197
Model 1&4.L.PN.F	4.2 [a]	63	5.1	176
Model 1&4.L.P.F	4.3 [a]	56	5.1	155
Model 5 class: *Stopping rule compensatory, decision rule compensatory—simple models*				
Model 5.1.PN	40.7 [a]	259	32.4	389
Model 5.1.P	0	193	9.4 [b]	286
Model 5.2.PN	48.1[a]	304	39.5	395
Model 5.2.P	0	352	9.4 [b]	431
Model 5.3.PN	0	357	9.4 [b]	449
Model 5.3.P	0	379	9.4 [b]	469
Model 1&5 class: *Stopping rule noncompensatory and compensatory, decision rule noncompensatory and compensatory—race models*				
Model 1&5.1.PN	15 [a]	243	8.3	372
Model 1&5.1.P	0	209	9.4 [b]	326
Model 1&5.1.PN.F	15.1 [a]	244	8.1	370
Model 1&5.1.P.F	0	205	9.4 [b]	324
Model 1&5.2.PN	8.1 [a]	133	6.8	248
Model 1&5.2.P	0	118	9.4 [b]	220
Model 1&5.2.PN.F	5.7 [a]	106	6.8	212
Model 1&5.2.P.F	0	96	9.4 [b]	189
Model 1&5.3.PN	0	109	9.4 [b]	223
Model 1&5.3.P	0	97	9.4 [b]	203
Model 1&5.3.PN.F	0	75	9.4 [b]	188
Model 1&5.3.P.F	0	70	9.4 [b]	167

Note. PN = Positive and negative cues. P = Positive cues. F = Forgetting cues. For decisions, RMSDs were calculated on the mean percentage of choices for the recognized city. For models that always decide for the recognized city, RMSDs for decisions will–by definition–always be 0 in the recognition group. For decision times, RMSDs were calculated on the median and the 1st and 3rd quartile and then averaged. Evaluations of the models' fit based on RMSDs should be complemented by visual inspections of the data produced by the models (see Figures 5.5-5.8 and Supplementary Online Material C: Figures 5.C1-5.C18). [a] These models do by definition not fit the decision of the recognition group, because they sometimes decide for the unrecognized city whereas participants in the recognition group always decide for the recognized city. [b] These models do by definition not fit the decision of the cue group, because they always decide for the recognized city whereas participants in the cue group sometimes decide for the unrecognized city.

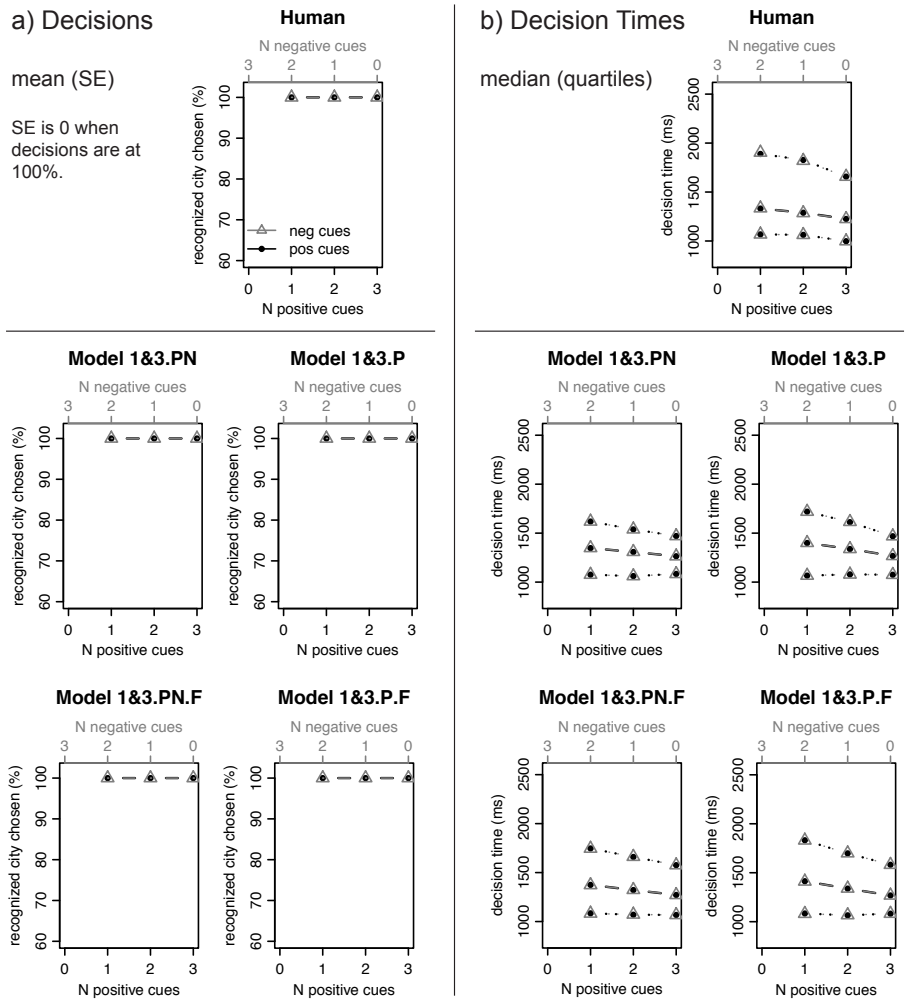


Figure 5.5 Decisions (a) and decision times (b) for the recognition group in Experiment 1. Human data and fits of the four models from the Model 1&3 class. Models are ordered from the top left to the bottom right in the same order as in Tables 5.2 - 5.5. In each graph, the upper grey x-axis shows the number of negative cues; the corresponding data points (decisions in Panel A, decision times in Panel B) are plotted in grey font (triangles). In each graph, the lower black x-axis shows the number of positive cues; the corresponding data points are plotted in black font (circles). For instance, in Panel B the median of the human decision times is 1335 ms for two negative cues and 1332 ms for one positive cue.

Figure 5.6 Decisions (a) and decision times (b) for the recognition group in Experiment 1. Human data and fits of those six models from the Model 1&5.2 and 1&5.3 classes that always decide for the recognized city in Experiment 1. Models are ordered from the top left to the bottom right in the same order as in Tables 5.2 - 5.5. In each graph, the upper grey x-axis shows the number of negative cues; the corresponding data points (decisions in Panel A, decision times in Panel B) are plotted in grey font (triangles). In each graph, the lower black x-axis shows the number of positive cues; the corresponding data points are plotted in black font (circles).

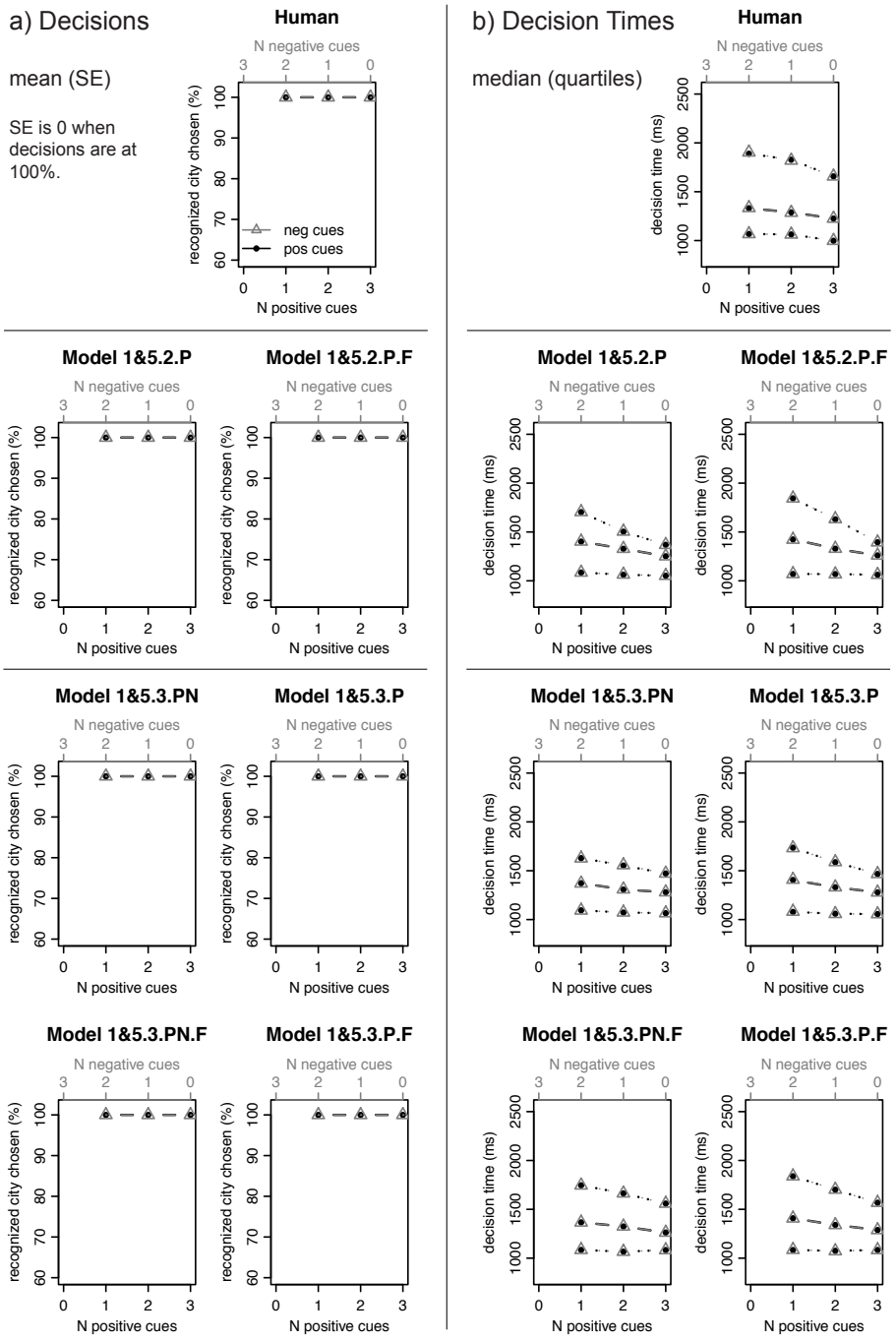
a) Decisions

mean (SE)

SE is 0 when decisions are at 100%.

b) Decision Times

median (quartiles)



Model 1&5 class that assume a decision criterion of 2 positive cues (Model 1&5.2.P, 1&5.2.P.F) fit the decisions and decision times well.

Importantly, while *technically* (i.e., by virtue of their RMSDs) Models 1&3.P.F and 1&5.3.P.F are the best-fitting models in Experiment 1's recognition group, all models of the 1&3 and 1&5.3 classes, as well as the P versions of the Model 1&5.2 class produce relatively similar fits. Therefore, we caution to declare any specific model from these classes to be considered the single winner. Rather, we would prefer to consider these classes the winner. In short, in Experiment 1's recognition group, the best-fitting model classes implement a race between Model 1's recognition-based noncompensatory stopping and decision rules and other processes; namely (i) Model 3's compensatory stopping rule and its recognition-based noncompensatory decision rule (i.e., as in the Model 1&3 class) as well as (ii) Model 5's compensatory stopping and decision rules (i.e., as in the Model 1&5 class).

We would like to add three observations with respect to the Model 1&5 class. First, note that Model 1&5.3.PN's and Model 1&5.3.PN.F's comparatively good fit of the recognition group's decisions (Figure 5.6) can be explained by Experiment 1's design. These two models need to retrieve 3 negative cues to decide against the recognized city ($D = 3$). As 3 negative cues were not taught in Experiment 1 (Table 5.1), Model 1&5.3.PN and Model 1&5.3.PN.F could not reach this decision criterion in Experiment 1, resulting in the models to always decide in favor of recognized cities. Had 3 negative cues been taught in Experiment 1, Model 1&5.3.PN and Model 1&5.3.PN.F would have produced decisions in favor of unrecognized cities, resulting in poor fits in the recognition group.[12]

Second, while one could thus argue that Model 1&5.3.PN's and Model 1&5.3.PN.F's good fit is an artifact of Experiment 1's design, the comparatively good fit of Model 1&5.3.P and Model 1&5.3.P.F is no such artifact: As these two models do not use negative cue knowledge, they never decide against unrecognized cities, but use recognition and positive cues to decide in favor of recognized ones. On the other hand, one may wonder whether compensatory, cue-based models that can *never* decide against unrecognized objects are theoretically plausible, or, what such models would add beyond models with simpler recognition-based, noncompensatory decision rules (e.g., as implemented by the Model 1&3 class).

Third, also Models 1&5.1.P and 1&5.1.P.F, which assume a decision criterion D of 1 positive cue, exhibit relatively small RMSDs (Table 5.4). By this token, also these representatives of the Model 1&5 class may belong to the winners. However, note that Models 1&5.1.P and 1&5.1.P.F produce a much smaller spread in the decision time distribution than the spread that can be found in the human times (Figure 5.C13 in Supplementary Online Material C).

[12] *To compare, the PN versions of the Model 1&5.1 and 1&5.2 classes (i.e., Model 1&5.1.PN, 1&5.1.PN.F, Model 1&5.2.PN, and 1&5.2.PN.F), do reach their decision criterion of* D = *1 and* D = *2 negative cues, respectively, letting these models occasionally decide for unrecognized cities. As a result, the PN versions of the Model 1&5.1 and 1&5.2 classes cannot fit the decisions in the recognition group (Table 5.4; Supplementary Online Material C, Figures 5.C13 and 5.C15).*

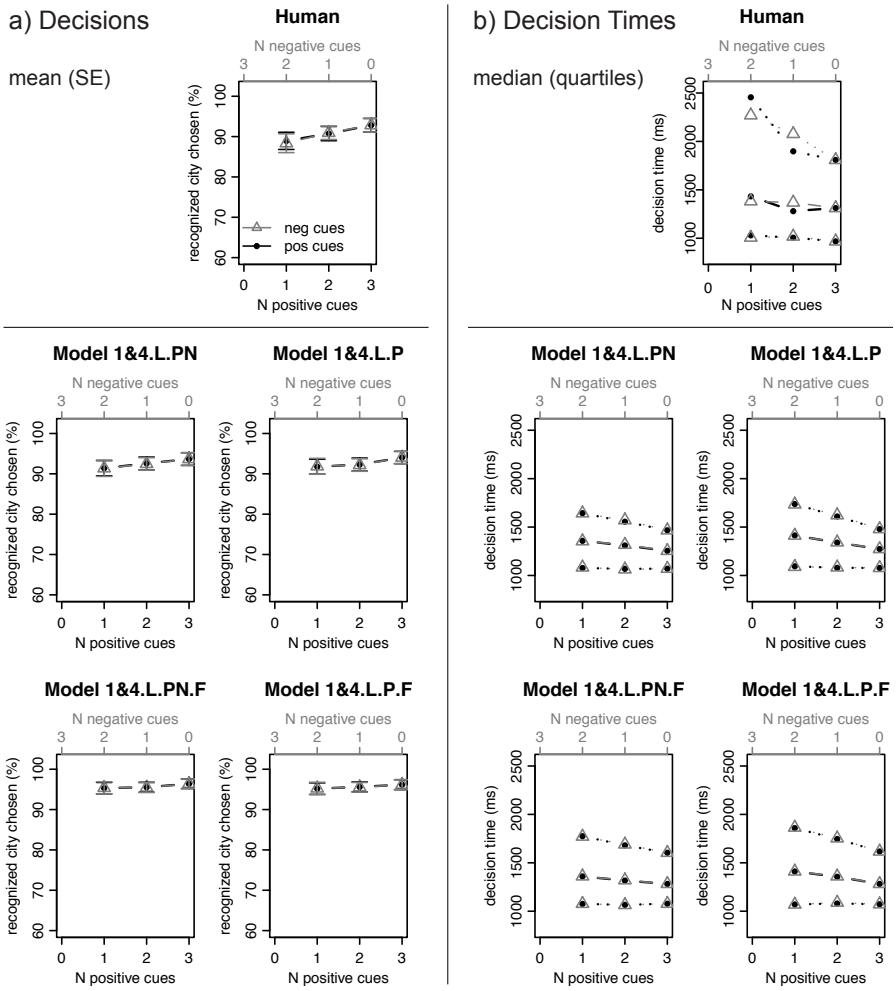


Figure 5.7 Decisions (A) and decision times (B) for the cue group in Experiment 1. Human data and fits of the four models from the Model 1&4.L class. Models are ordered from the top left to the bottom right in the same order as in Tables 5.2 - 5.5. In each graph, the upper grey x-axis shows the number of negative cues; the corresponding data points (decisions in Panel A, decision times in Panel B) are plotted in grey font (triangles). In each graph, the lower black x-axis shows the number of positive cues; the corresponding data points are plotted in black font (circles). For instance, in Panel A the mean percentage of participants' choices for the recognized city is 88 for two negative cues and 89 for one positive cue.

Cue group

Figure 5.7 shows the human decisions and decision times as well as the decisions and decision times produced by the Model 1&4.L class, which is the class that best fits the combination of decisions and decision times in the cue group. As can be seen, the human decisions and decision times as well as the models' decisions and decision times

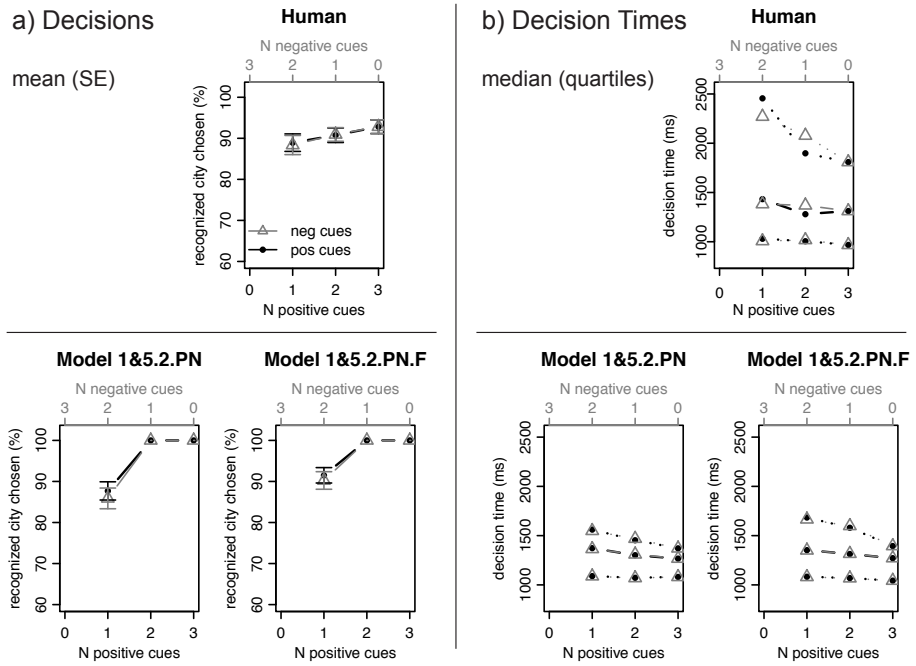


Figure 5.8 Decisions (A) and decision times (B) for the cue group in Experiment 1. Human data and fits of those two models from the Model 1&5.2 class that sometimes decide against the recognized city in Experiment 1. Models are ordered from left to right in the same order as in Tables 5.2 - 5.5. In each graph, the upper grey x-axis shows the number of negative cues; the corresponding data points (decisions in Panel A, decision times in Panel B) are plotted in grey font (triangles). In each graph, the lower black x-axis shows the number of positive cues; the corresponding data points are plotted in black font (circles).

vary as a function of cues. The decision times show a large spread. While the Model 1&4.L class emerges as the best-fitting class, it is difficult to rank order the models *within* that class in terms of their RMSDs. As Table 5.4 shows, Model 1&4.L.P.F fits the decision times best; however, this model does not produce the smallest RMSDs for the decisions, which are produced by Model 1&4.L.PN.

Let us turn to a couple of other models that may, perhaps, be considered to belong to the winners in the cue group. First, as can be seen in Table 5.4 (and Figure 5.C8 in Supplementary Online Material C), the Model 1&4.H class, which differs from the Model 1&4.L class only in the base level activation of the big chunk, produces a good fit of the decision times, while not fitting the decisions as well as the 1&4.L class. Second, Table 5.4 suggests that also the PN versions of the Model 1&5.2 class (i.e., Model 1&5.2.PN, 1&5.2.PN.F) produce a relatively good fit to the cue group's combination of decisions and decision times. However, as a visual inspection of Figure 5.8 reveals, these models produce an abrupt drop in decisions for the recognized city as soon as the decision criterion of $D = 2$ negative cues is reached. The human data do not exhibit such a drop. Much the same can be said with respect to the PN versions of

Table 5.5 Root mean square deviations between the model and the human data in Experiment 2.

	Recognition group		Cue group	
	Decisions (%)	Decision times (ms)	Decisions (%)	Decision times (ms)
Model 1 class: *Stopping and decision rules noncompensatory—simple model*				
Model 1	0	279	15.9 [b]	498
Model 2 class: *Stopping rule compensatory, decision rule noncompensatory—simple models*				
Model 2.PN	0	255	15.9 [b]	290
Model 2.P	0	307	15.9 [b]	320
Model 3 class: *Stopping rule compensatory, decision rule noncompensatory—simple models*				
Model 3.PN	0	454	15.9 [b]	380
Model 3.P	0	531	15.9 [b]	410
Model 1&3 class: *Stopping rule noncompensatory and compensatory, decision rule noncompensatory—race models*				
Model 1&3.PN	0	101	15.9 [b]	170
Model 1&3.P	0	135	15.9 [b]	145
Model 1&3.PN.F	0	134	15.9 [b]	131
Model 1&3.P.F	0	179	15.9 [b]	109
Model 4 class: *Stopping rule compensatory, decision rule compensatory—simple models*				
Model 4.H.PN	11.7 [a]	590	6.8	435
Model 4.H.P	12 [a]	666	6.6	474
Model 4.L.PN	57.8 [a]	623	45.1	456
Model 4.L.P	57.1 [a]	699	44.9	500
Model 1&4 class: *Stopping rule noncompensatory and compensatory, decision rule noncompensatory and compensatory—race models*				
Model 1&4.H.PN	1.6 [a]	100	14.4	164
Model 1&4.H.P	1.6 [a]	151	14.5	140
Model 1&4.H.PN.F	0.9 [a]	138	15.1	124
Model 1&4.H.P.F	0.8 [a]	180	15.1	88
Model 1&4.L.PN	7.4 [a]	115	10.9	166
Model 1&4.L.P	7.2 [a]	150	11.2	145
Model 1&4.L.PN.F	4.2 [a]	147	13.1	125
Model 1&4.L.P.F	4 [a]	204	12.9	95
Model 5 class: *Stopping rule compensatory, decision rule compensatory—simple models*				
Model 5.1.PN	60.1 [a]	193	44.8	331
Model 5.1.P	0	436	15.9 [b]	409
Model 5.2.PN	56.7 [a]	284	40.8	295
Model 5.2.P	0	472	15.9 [b]	373
Model 5.3.PN	44.7 [a]	453	28.5	380
Model 5.3.P	0	531	15.9 [b]	410
Model 1&5 class: *Stopping rule noncompensatory and compensatory, decision rule noncompensatory and compensatory—race models*				
Model 1&5.1.PN	22 [a]	166	7.2	321
Model 1&5.1.P	0	177	15.9 [b]	251
Model 1&5.1.PN.F	21.8 [a]	163	6.9	320
Model 1&5.1.P.F	0	208	15.9 [b]	236
Model 1&5.2.PN	13 [a]	123	2.9	206
Model 1&5.2.P	0	142	15.9 [b]	167
Model 1&5.2.PN.F	10.5 [a]	106	5.6	185
Model 1&5.2.P.F	0	175	15.9 [b]	131
Model 1&5.3.PN	5.8 [a]	99	10.2	162
Model 1&5.3.P	0	146	15.9 [b]	146
Model 1&5.3.PN.F	3.1 [a]	131	12.6	135
Model 1&5.3.P.F	0	167	15.9 [b]	104

Note. PN = Positive and negative cues. P = Positive cues. F = Forgetting cues. For decisions, RMSDs were calculated on the mean percentage of choices for the recognized city. For models that always decide for the recognized city, RMSDs for decisions will–by definition–always be 0 in the recognition group. For decision times, RMSDs were calculated on the median and the 1st and 3rd quartile and then averaged. Evaluations of the models' fit based on RMSDs should be complemented by visual inspections of the data produced by the models (see Figures 5.9-5.12 and Supplementary Online Material C: Figures 5.C19-5.C36). [a] These models do by definition not fit the decision of the recognition group, because they sometimes decide for the unrecognized city whereas participants in the recognition group always decide for the recognized city. [b] These models do by definition not fit the decision of the cue group, because they always decide for the recognized city whereas participants in the cue group sometimes decide for the unrecognized city.

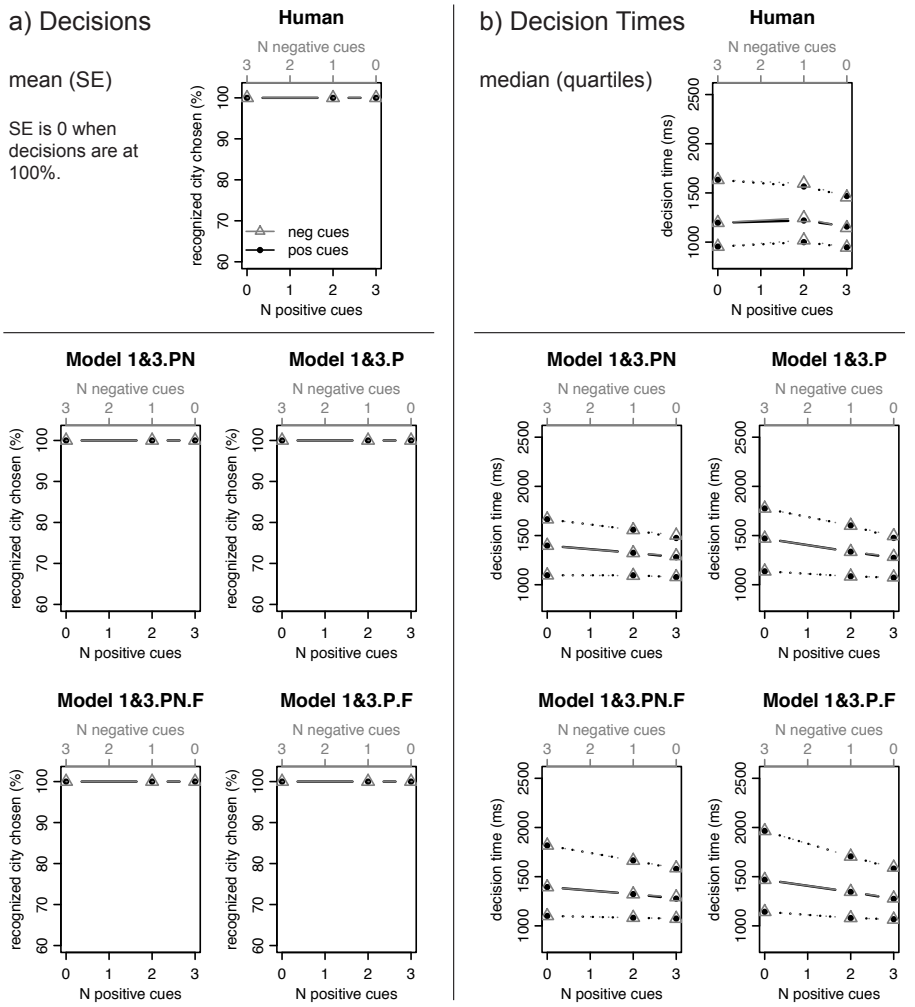


Figure 5.9 Decisions (A) and decision times (B) for the recognition group in Experiment 2. Human data and predictions of the four models from the Model 1&3 class. Models are ordered from the top left to the bottom right in the same order as in Tables 5.2 - 5.5. In each graph, the upper grey x-axis shows the number of negative cues; the corresponding data points (decisions in Panel A, decision times in Panel B) are plotted in grey font (triangles). In each graph, the lower black x-axis shows the number of positive cues; the corresponding data points are plotted in black font (circles).

the Model 1&5.1 class (Figure 5.C13 in Supplementary Online Material C), which produce an even steeper drop in the decisions, and which fit the spread of the decision times less well than the Model 1&5.2 class.

In short, the cue group's best-fitting models are members of the Model 1&4.L class. This model class implements a race between Model 1's noncompensatory stopping rule and Model 4's compensatory stopping rule as well as a race between Model 1's

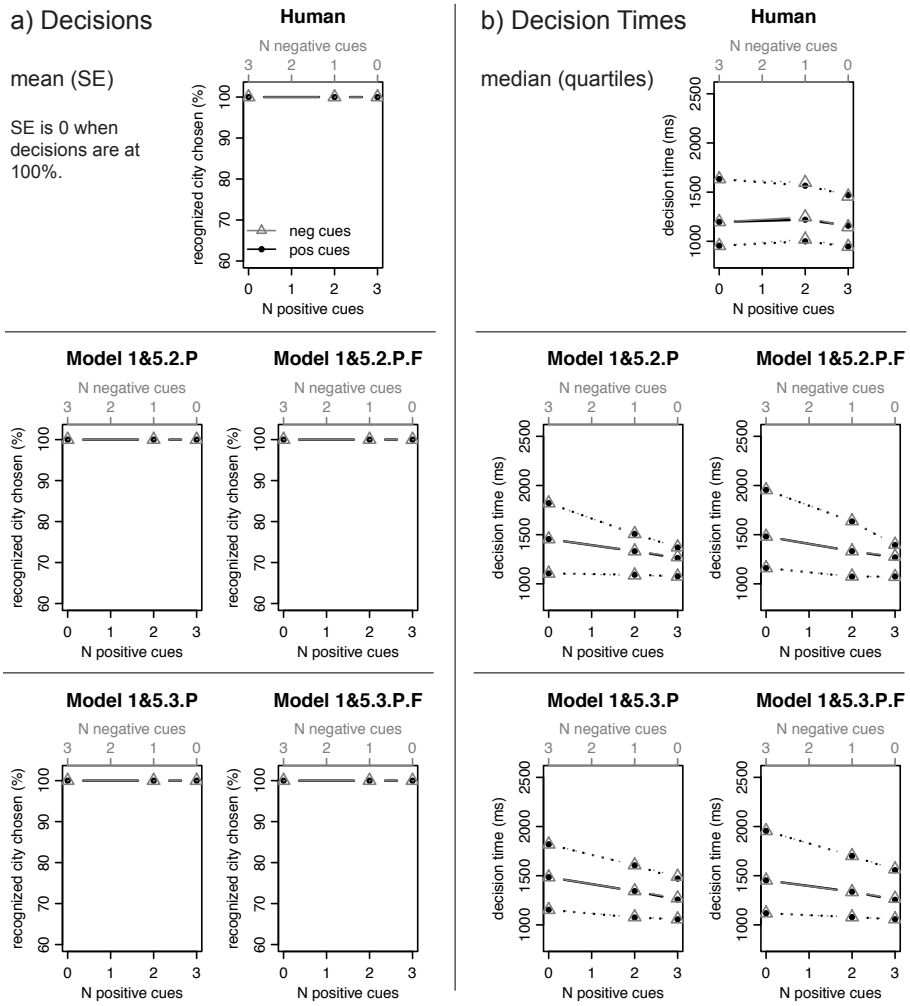


Figure 5.10 Decisions (A) and decision times (B) for the recognition group in Experiment 2. Human data and predictions of those four models from the Model 1&5.2 and 1&5.3 classes that always decide for the recognized city in Experiment 2. Models are ordered from the top left to the bottom right in the same order as in Tables 5.2 - 5.5. In each graph, the upper grey x-axis shows the number of negative cues; the corresponding data points (decisions in Panel A, decision times in Panel B) are plotted in grey font (triangles). In each graph, the lower black x-axis shows the number of positive cues; the corresponding data points are plotted in black font (circles).

noncompensatory decision rule and Model 4's compensatory decision rule, assuming implicit, intuitive knowledge about the cities' sizes to be responsible for occasional decisions in favor of unrecognized cities.

Results of the Model Generalization Competition in Experiment 2

To test how well these results generalize to another data set, we let all 39 models predict the human decisions and decision times from Experiment 2. In doing so, we populated the models' declarative memory with each individual participant's recognition and cue knowledge, using participants' responses in the recognition task and cue-memory task of Experiment 2 —just as we did in Experiment 1. And as in Experiment 1, we ran the models on the trials of each individual participant in the decision task of Experiment 2. Following our principle of predictive modeling, we kept all models' production rules as well as the values of all models' parameters identical to those used in Experiment 1. Table 5.5 summarizes the results for all models. In what follows, we will mainly discuss those models that generalized best (for all other models' generalizability and a complete set of graphs of all models' predictions see Supplementary Online Material C.)

Recognition group

Figures 5.9 and 5.10 show the human decisions and decision times as well as the corresponding data produced by the best-generalizing models in the recognition group. These are representatives of the Model 1&3 class, as well as those representatives from the Model 1&5 class that assume a decision criterion of 2 and 3 positive cues (Models 1&5.2.P, 1&5.2.P.F, 1&5.3.P, 1&5.3.P.F). As can be seen, all winning models correctly predict that decisions do not vary as a function of cues. The models also predict the overall pattern and spread of the decision times well. Importantly, as the RMSDs in Table 5.5 show, the technically best-generalizing model, Model 1&3.PN, belongs to the Model 1&3 class, which also was one of the winning model classes in Experiment 1, lending, perhaps, further support to the 1&3 class.

Note that also Model 1&5.1.P—and to a lesser extent Model 1&5.1.P.F—exhibit relatively small RMSD in Table 5.5. However, as in Experiment 1, these models fail to predict the spread of the human decision times (Figure 5.C31 in Supplementary Online Material C).

In short, for the recognition group, members of the Model 1&3 class are among the best models in both experiments. Also the versions of the Model 1&5.2 and 1&5.3 class that use only positive cues perform well in both experiments. The versions of the Model 1&5.3 class that use positive *and negative* cues fitted Experiment 1's recognition group well (cf. Figure 5.6), but do not predict the recognition group's decisions in Experiment 2. Recall that these two models need to retrieve 3 negative cues to decide against the recognized city. As 3 negative cues were not taught in Experiment 1 (cf. Table 5.1), the models did not reach their decision criterion, leading them to always decide in favor of recognized cities. In Experiment 2, in contrast, 3 negative cues *were* taught. Correspondingly, the models do reach their decision criterion, leading them to occasionally decide against the recognized city, this way mismatching the recognition group data. However, as we explain next, these models (Model 1&5.3.PN and Model 1&5.3.PN.F) turn out to generalize well to Experiment 2's cue group.

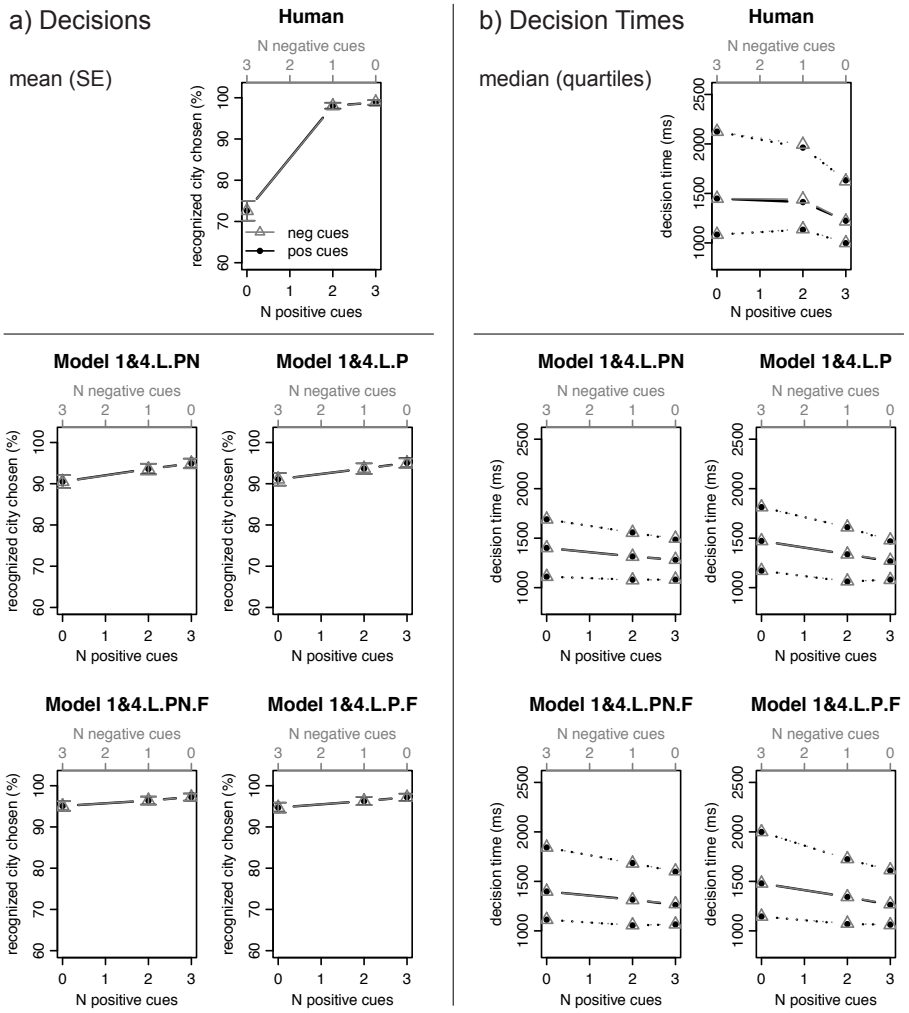


Figure 5.11 Decisions (A) and decision times (B) for the cue group in Experiment 2. Human data and predictions from the four models from the Model 1&4.L class. Models are ordered from the top left to the bottom right in the same order as in Tables 5.2 - 5.5. In each graph, the upper grey x-axis shows the number of negative cues; the corresponding data points (decisions in Panel A, decision times in Panel B) are plotted in grey font (triangles). In each graph the lower black, x-axis shows the number of positive cues; the corresponding data points are plotted in black font (circles).

Cue group

Figures 5.11 and 5.12 show the human data and the best-generalizing models in the cue group. These are the Model 1&4.L class as well as those representatives of the Model 1&5.2 and 1&5.3 classes that use positive and negative cues.

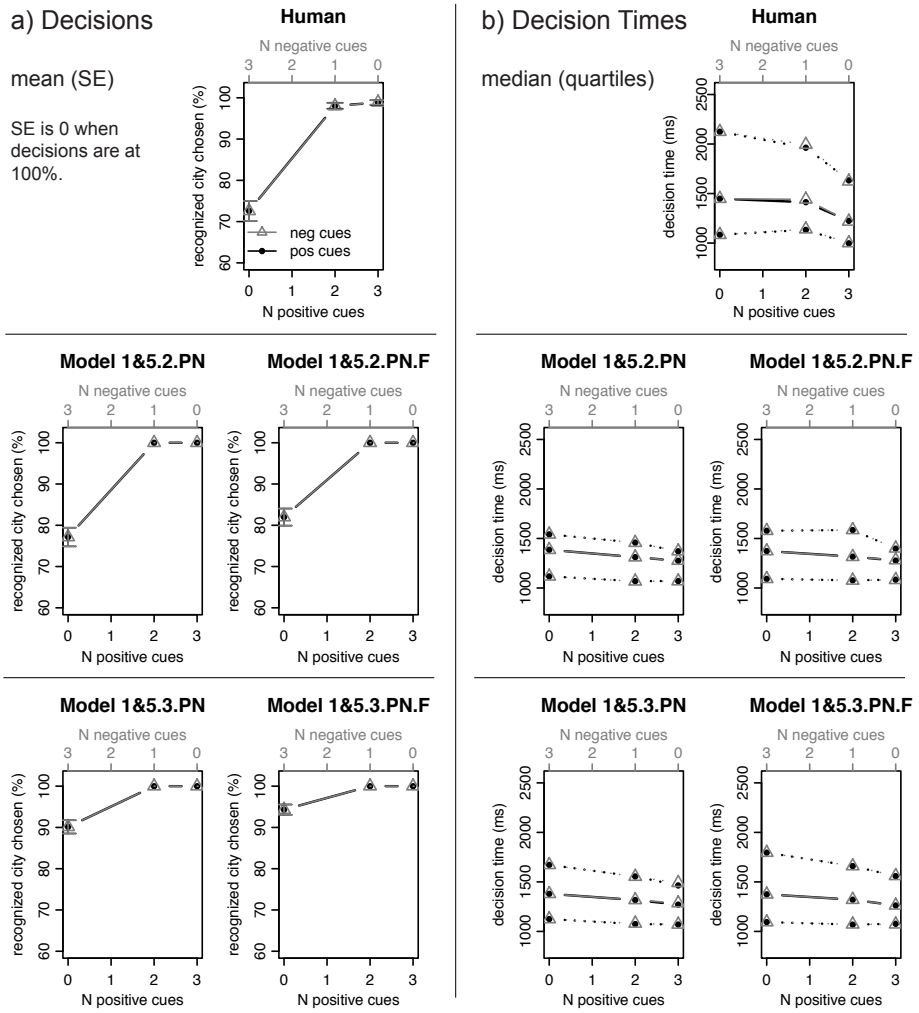


Figure 5.12 Decisions (A) and decision times (B) for the cue group in Experiment 2. Human data and predictions from those four models from the Model 1&5.2 and 1&5.3 classes that sometimes decide against the recognized city in Experiment 2. Models are ordered from the top left to the bottom right in the same order as in Tables 5.2 - 5.5. In each graph, the upper grey x-axis shows the number of negative cues; the corresponding data points (decisions in Panel A, decision times in Panel B) are plotted in grey font (triangles). In each graph the lower black, x-axis shows the number of positive cues; the corresponding data points are plotted in black font (circles).

Let us first turn to the decisions of the Model 1&4.L class, which fitted the data best in Experiment 1. As in Experiment 1, the human decisions, as well as the decisions of the models vary as a function of cues. However, in Experiment 2, the human decisions are strongly influenced by three negative cues (i.e., corresponding to zero positive cues). Having been adjusted to Experiment 1, in which participants were taught a maximum

of two negative cues (Table 5.1), the Model 1&4.L class fits the decisions for zero and one negative cue well, but has difficulties to predict the large effect of three negative cues in Experiment 2 (Figure 5.11). Much the same can be said with respect to the Model 1&4.H class, which, as in Experiment 1, does not predict the decisions as well as the 1&4.L class (Table 5.5; Figure 5.C26 in Supplementary Online Material C).

In contrast, consider the decisions of the PN versions of the Model 1&5.2 and 1&5.3 classes (Model 1&5.2.PN, 1&5.2.PN.F, 1&5.3.PN, 1&5.3.PN.F).). As shown in Figure 5.12, these models *do* predict a large effect of negative cues on the decisions once their decision criterion of *D* negative cues is reached. Models 1&5.2.PN and 1&5.2.PN.F, which decide against the recognized city as soon as two negative cues have been retrieved, predict the pattern in the human decisions best (Table 5.5, Figure 5.12).

Figures 5.11 and 5.12 also show the decision times. The models from the 1&4.L class as well as the PN versions of the 1&5.2, and 1&5.3 classes are able to approximate the human decision time pattern and its spread. However, Models 1&5.2.PN and 1&5.2.PN.F, which predict the decisions best, do not predict the decision times as well as the representatives of the 1&4.L class and the PN versions of the 1&5.3 class (Table 5.5), making it difficult to rank order the best model classes in terms of their performance.

Note that, as in Experiment 1, also the PN versions of the Model 1&5.1 class produce a drop in the decisions once its decision criterion of $D = 1$ negative cue is reached. However, this drop is steeper than in the human data and the model class fails to predict the spread of the decision times (Figure 5.C32 in Supplementary Online Material C.)

In short, the winning model classes in Experiment 2's cue group are essentially identical to those that won in Experiment 1's cue group—with two relevant caveats. First, in Experiment 2, besides the Model 1&4.L and 1&4.H classes, and the PN versions of the 1&5.2 classes, also the PN versions of the Model 1&5.3 class may be considered to belong to the winners. Second, in Experiment 1, the Model 1&4.L class fitted the decisions and decision times best. In Experiment 2, it is more difficult to establish a rank order of these classes' ability to predict the human data, as those models that predict the decisions best do not predict the decision times best.[13]

[13] *The results reported throughout this chapter are based on data that has been collapsed across participants. To explore whether the results hold when the data is not collapsed, we ran a second analysis. Using the very same model parameter values as the ones reported above, we calculated the RMSD between each participant and each model and then averaged the resulting RMSDs across participants. These averaged RMSDs were generally higher than the RMSDs calculated for the collapsed data, which is not surprising, as the models' parameter values were fitted to the collapsed data and not to the individual data. Importantly, overall the same model classes that won the model competition on the collapsed data emerged as the winning model classes also in this second, exploratory analysis. However, in several (but not all) cases* within *the winning model classes, the rank order of the models' goodness of fit changed. For instance, in our original analysis of the collapsed data of Experiment 1's recognition group, Model 1&3.P.F and Model 1&5.3.P.F were technically the best models. In the second analysis, Model 1&3.PN and Model 1&5.3.PN were the best models. At the same time, in Experiment 2's recognition group, in both, our original analysis on the collapsed data as well as in the second analysis, Model 1&3.PN fitted best. Importantly, the RMSD differences within the different Model classes are small in both analyses. This further suggests that the rank order within model classes should be interpreted with caution and supports the point that it is model classes, rather than single models that can be identified as winners in our model comparison (see, e.g., the result section on the best fitting models in the recognition group of Experiment 1).*

General Discussion

Much research has investigated how people make decisions based on a sense of the accessibility of memories, as assumed by the recognition heuristic and related models (Bruner, 1957; Jacoby & Dallas, 1981; Pachur et al., 2011; Pohl, 2011; Tversky & Kahneman, 1973). At the same time, in the field of accessibility-based decision making and beyond, many have criticized the lack of specification of process hypotheses (e.g., Dougherty et al., 2008; Dougherty et al., 1999; Gigerenzer, 1996, 1998; Keren & Schul, 2009; A. Newell, 1973). Particularly the recognition heuristic has triggered a controversy about what processes describe people's decisions best when they make inferences from the accessibility of memories: Do people rely on this noncompensatory heuristic, ignoring further knowledge, or do they use compensatory strategies instead?

In this chapter, we provided a primer on how the precision of corresponding process hypotheses can be increased. Using the ACT-R cognitive architecture, we specified process hypotheses about accessibility-based decisions in 39 quantitative process models. These models do not only capture decision processes, but also the interplay of decision processes with perceptual, memory, intentional, and motor processes. Moreover, by implementing a number of decision models that had originally been defined at different levels of description into *one* architectural modeling framework, we made these models comparable, providing a basis for detailed, multi-experiment model comparisons to be conducted in future research. Finally, we conducted a first model comparison ourselves, re-analyzing two previously published data sets.

Even though the main objective of this model comparison was to illustrate how such comparisons can be conducted rather than to conclusively identify the best model, in what follows we will first discuss our model comparison's results. We will close by turning to a number of broader methodological issues.

Dissolving Dichotomies by Implementing More Than One Process: Race Models

Both in fitting existing data and in generalizing to new data, representatives of the race model classes performed best in our model competition. As such, the winners are models that implement recognition-based noncompensatory processes side by side with cue-based compensatory ones, suggesting that in one part of the trials in the decision task noncompensatory processes governed information retrieval and/or decision making, while in the other part compensatory processes were dominant. Specifically, our results highlight the possibility that even people who always responded with recognized cities (i.e., as in the recognition group) most likely retrieved and encoded cues in at least some of the trials. People who sometimes responded with unrecognized cities (i.e., as in the cue group), in turn, most likely based their decisions on cues in some of the trials but ignored these cues and relied on recognition in others. These results dissolve the dichotomy between cue-based compensatory and recognition-based noncompensatory processes that is often assumed in the literature and that has fuelled debates about

the recognition heuristic (e.g., Pohl, 2006, 2011; Richter & Späth, 2006, see above). Moreover, these results cast, perhaps, some doubt on a simplifying assumption that is central to this debate: By classifying a person exclusively as either a noncompensatory or a compensatory decision maker, previous studies had (at least implicitly) assumed that a person's decision processes do not vary across the trials of a decision task (e.g., Glöckner & Bröder, 2011; Marewski, Gaissmaier, Schooler et al., 2010).[14]

We hasten to add that our analyses entailed collapsing the data across participants' responses, which severely limits the possibility to draw conclusions about individual persons' decision processes. We suggest for future research to tackle this question, by using more exhaustive human data sets and analyses.

Models Implementing One Decision Process: Simple Models

Models that implement merely one type of decision process, namely noncompensatory or compensatory, did not account as well for people's behavior as the winning race models. Let us first turn to the noncompensatory models, and then to the compensatory ones.

Noncompensatory models

The strictly noncompensatory Model 1, which neither retrieves nor uses cues for decisions, did not accurately predict participants' decision times, even for participants who always chose the recognized city (Supplementary Online Material C, Figures 5.C1, 5.C19). As such, our results cast doubts on recognition heuristic implementations that assume noncompensatory recognition-based stopping and decision rules. Much the same can be said with respect to those recognition heuristic implementations that retrieve cues but do not use them for decisions: Also the Model 2 and 3 classes, which implement corresponding cue-based compensatory stopping and recognition-based noncompensatory decision rules, did not account well for people's behavior (Supplementary Online Material C, Figures 5.C1, 5.C19). However, the relative success of the 1&3 race Model class lends support to a *combination* of both recognition heuristic implementations: As the Model 1&3 class includes Model 1 *and* Model 3 as components, our results suggest that a combination of these two recognition heuristic implementations may reflect people's decision processes in the comparisons of cities (Gigerenzer & Goldstein, 2011).

We would like to add two points. First, while representatives of the Model 1&3 class are both among Experiment 1's best fitting and among Experiment 2's best generalizing models, also those representatives of the 1&5 Model class that rely on positive cues in addition to recognition were able to account for behavior well. This result leads us to stress that it may be similarly plausible for noncompensatory, recognition-based stopping and decision rules to govern a part of the comparisons of

[14] *The approach to classify a person either exclusively as a compensatory decision maker or as a noncompensatory one is also common in studies on people's use of other heuristics, such as take-the-best (Bröder, 2003; Bröder & Gaissmaier, 2007; Bröder & Schiffer, 2003, 2006).*

two cities (i.e., Model 1), while compensatory, cue-based processes govern the other part (i.e., Model 5). On the other hand, the Model 1&3 class provides, arguably, a more parsimonious explanation for the human data than the Model 1&5 class.

Second, we implemented just one *strictly* noncompensatory variant of the recognition heuristic: Model 1, which has *both* a recognition-based noncompensatory stopping *and* decision rule. It is to be expected that pitting this single strictly noncompensatory model against a total of 38 other models may have biased the outcome of the model comparison against strictly noncompensatory models.

Compensatory models

We implemented two types of strictly compensatory models. In assuming that subsymbolic pathways and spreading activation give rise to implicit, intuitive knowledge that governs compensatory decision processes, the Model 4 class implements a central feature of Glöckner and Betsch's (2008) parallel constraint satisfaction model. The parallel constraint satisfaction model has been argued to account for behavior better than the recognition heuristic—at times without the model having been applied to data (e.g., Hilbig & Pohl, 2009; Hochman et al., 2010); see Glöckner and Bröder (2011) for a test that does apply the model to data.

The Model 5 class assumes symbolic pathways to be responsible for compensatory processes, and as such, decisions to be based on explicit, deliberate knowledge. Also models from this class have been discussed as antipodes to the recognition heuristic, almost always with such models not being applied to data (e.g., Hilbig & Pohl, 2009; B. R. Newell & Shanks, 2004; Oeusoonthornwattana & Shanks, 2010; Pohl, 2006; Richter & Späth, 2006), or with the models having been applied to data, but without using the models to quantitatively predict decision times (Marewski et al., 2009; Pachur & Biele, 2007).

Whereas both Model 4 and Model 5 classes were able to account for some aspect of the human data in the cue group, neither turned out to be sufficient (Supplementary Online Material C; Figures 5.C6, 5.C12, 5.C24, 5.C30). Instead, the race models of the Model 1& 4 class, that is, combinations of the implicit, intuitive processes assumed by Model 4 and the noncompensatory, recognition-based processes of Model 1 were able to fit participants' data best in Experiment 1. In Experiment 2, race models of the 1&4 class were also among the best-generalizing models; however, here representatives of the Model 1&5 class rivaled their performance. In short, with respect to strictly compensatory models, the current data suggest that the simple Model 4 and 5 classes are insufficient.

Methodological Considerations

Model specification

At the close of this chapter, we would like to stress five points. First, most of the hypotheses about accessibility-based decisions tested here had only been formulated

verbally in the literature. As a result, the outcomes of our model comparison also depend on our choices of how to implement such verbal hypotheses into detailed computational models in ACT-R. That is, we cannot rule out the possibility that different implementations will lead to different results in model competitions. It is important to realize, however, that this *specification problem* (Lewandowsky, 1993), namely, how to translate an underspecified hypothesis into a detailed model, is not a problem specific to research on accessibility-based decisions, but can also emerge when using cognitive architectures to implement hypotheses about cognitive processes in other areas of research, including when implementing classic decision strategies such as *elimination-by-aspects* (Tversky, 1972). Here we dealt with this problem by following the principles of competitive and nested modeling, leading us to implement a large number of variants of the accessibility-based strategies discussed in the literature.

Architecture

Second, the lack of specification that many decision strategies exhibit is also problematic for another reason: Often it is not clear what drives a strategy's ability to account for process data. Is it an unspecified assumption, for example about memory, perceptual, or motor processes? Or is it the decision strategy itself that carries the burden of explanation? As A. Newell (1990) puts it, a theory that deals with only one component of behavior (e.g., decision making) while ignoring the rest (e.g., memory) "flirts with trouble from the start" (p. 17). In our view, models of decision making should therefore be specified at an architectural level, spelling out not only decision processes, but also how these processes interweave with other cognitive processes.

Modeling principles

Third, we deem the two experimental data sets and analyses reported here to be insufficient to conclusively identify the best process model. For instance, as discussed above, some of our 39 models' ability to account for the experimental data was similar. However, we would like to point out that we were able to obtain a more differentiated picture of the models' performance than one might have expected, given how large the number of tested models was. We attribute those relatively clear-cut results of our model competition to the five methodological principles we embraced. For instance, had we just fitted median decision times and not additionally let the models fit and predict the decision times' 1st and 3rd quartiles, then it would have been more difficult to judge which models account for decision times best, because different models may be able to produce similar median times, but different spreads for the underlying decision time distributions. Similarly, had we not constrained the models by estimating recognition and retrieval parameters from separate recognition and cue retrieval tasks and then keeping all parameters constant across all models, it might have been more difficult to tell whether a failure of a model to account for decision times should be attributed to the model's assumptions about recognition and retrieval processes or to the model's assumptions about decision processes.

Strategy selection

Fourth, we would like to point out that comparative tests of process models of decision strategies such as the ones we conducted above are incomplete if they are not informed by theories of strategy selection. Such theories predict in what situations and tasks a given decision strategy will be relied upon and in what situations and tasks a strategy will not come into play (Busemeyer & Myung, 1992; Lovett & Anderson, 1996; Marewski & Schooler, 2011; Rieskamp & Otto, 2006). Without such a theory, rejecting a model of decision making simply because it does not predict behavior well in a certain situation or task is problematic. There are at least two potential reasons why a decision strategy does not predict behavior. One is (a) that the strategy *per se* is generally not a good model of behavior. An alternative reason is (b) that the decision strategy is not relied upon, because people (or the corresponding selection mechanisms) *choose* not to use it in a particular situation. For instance, in the cue group of Experiment 1, Models of the 1&4.L class fitted decisions and decision times best, lending support to an implicit use of cue knowledge. In Experiment 2, results were different. Whereas also in this experiment, Models of the 1&4.L class predicted the human decisions well for zero and one negative cues, models assuming more deliberate, explicit decision processes (i.e., Models of the 1&5.2 class) turned out to be the better predictors for decisions when three negative cues were known about the recognized city. The fact that the Model 1&4.L's class relative success did not completely generalize from Experiment 1 to Experiment 2 could not only be interpreted as (a) challenging the validity of this model class, but also as (b) the difference in the design of the two experiments (Table 5.1) having resulted in a change in the decision strategies participants employed. A model of strategy selection that predicts when a given decision strategy will be used (and when not) could help to establish which of these two interpretations is likely to represent the better one.

Generalizability across experimental paradigms

Fifth, we would also like to stress that different experimental paradigms can require specifying different cognitive processes in the same decision model. Pachur et al.'s (2008) Experiment 1 and 2, which we re-analyzed here for the purpose of illustrating our 39 ACT-R models, entailed teaching participants cue knowledge about the cities (e.g., whether a city has an airport). It is not clear to what extent the results of our model comparison will generalize to experiments where participants have acquired their cue knowledge naturally, that is, is outside of the laboratory. For instance, in teaching the cue knowledge in Pachur et al.'s experiments, all to-be-learned cues were presented with equal frequency, making it likely that all cues exhibit similar base level activation in memory and have similar probabilities and speeds of retrieval. In experiments where knowledge is acquired naturally, the activation of different pieces of information will vary as a function of the environment, which can result in different probabilities and speed of retrieval for different pieces of information (see Marewski & Schooler, 2011, for corresponding ACT-R modeling efforts). In such experiments,

different decision strategies may emerge as the winners than those we identified in our model comparisons. We encourage future research to tackle this question, because experimental paradigms involving naturally acquired information may be considered an ideal test-bed for the recognition heuristic (Gigerenzer & Goldstein, 2011; Pachur et al., 2008).

Conclusion: Beyond Qualitative Hypotheses and Simplifying Dichotomies

"Psychology […] attempts to conceptualize what it is doing […] How do we do that? Mostly […] by the construction of oppositions—usually binary ones. We worry about nature versus nurture, about central versus parallel, and so on." These lines written by Allen Newell in 1973 (p. 287) still reflect much research in the decision sciences today that centers on dichotomies such as compensatory versus noncompensatory processes. Also much of contemporary research on accessibility-based decisions and on the recognition heuristic suffers from this state of affairs (Tomlinson, Marewski, & Dougherty, 2011). By developing models of accessibility-based decisions within an architecture, we have taken a small step toward replacing such dichotomies and the qualitative processes hypotheses associated with them, with detailed, quantitative models (see also, Anderson, 2007; Dougherty et al., 1999; Marewski & Schooler, 2011; Nellen, 2003; A. Newell, 1990; Schooler & Hertwig, 2005).

To conclude, we would like to highlight that often there may exist many different models, all of which are equally capable of reproducing and explaining data—a dilemma that is also known as the *identification problem* (Anderson, 1976). As a result it appears unreasonable to ask which of many process models is more "truthful"; rather, one needs to ask which model is better than another given a set of criteria, for example, the models' degree of specification or its generalizability to new tasks. As Box (1979) puts it - and we agree - "All models are wrong, but some are useful" (p. 202). Importantly, however, while many functionally equivalent models may exist, there are infinite numbers of underspecified models for which nobody will ever be able to decide whether one is better than another, given a set of criteria. Thus, even though all models may be wrong, often there is no better alternative than making them as precise as possible.

Summary & Conclusion

In which I summarize the findings and discuss why precision matters.

Chapter 9

Summary & Conclusion

Summary & Conclusion

Newell (1973) closed his 20 questions paper with the words: "Maybe all is well, […], and when we arrive in 1992 […] we will have homed in to the essential structure of the mind." (p. 306). Have we, now in 2011, reached this goal? Undeniably, progress has been made since the 70s. The use of computational models has increased the precision of theoretical predictions and cognitive architectures like ACT-R provide a promising tool towards understanding "the essential structure of the mind". Yet, underspecified verbal models and simplifying dichotomies still enjoy great popularity.

Currently, a vast amount of research centers on the debate of whether reasoning and decision making is based on implicit, automatic, and high-capacity, or on explicit, deliberate, and low-capacity processes (e.g., Dijksterhuis, Bos, Nordgren, & Van Baaren, 2006). Another popular dichotomy can be found in the decision sciences, where there is an ongoing discussion whether decisions can better be described by simple non-compensatory heuristics or by more complex compensatory strategies (see e.g., the special issues by Marewski, Pohl, et al., 2010; Marewski, et al., 2011a, 2011b). While tests of binary oppositions can certainly lead to interesting insights, they are, as Newell warned, not always useful. Often, the apparently opposing aspects even represent "two sides of the same coin". The real challenge for understanding cognition lies, therefore, not so much in testing opposing aspects against each other, but in understanding their respective contribution and interaction. To do so, we need to understand the underlying cognitive processes. In this thesis I have shown how the precision provided by computational models can help us to meet this challenge.

Memory Activation in Diagnostic Reasoning

The starting point of this dissertation was the idea that automatic memory processes can facilitate diagnostic reasoning by providing the reasoner with an adaptive subset of potential hypotheses from memory. The work presented in Chapters 2, 3, and 4 not only supports this idea, but also identifies and tests potential underlying memory mechanisms.

In Chapters 2 and 3 we tested whether and how observed symptoms can activate associated explanations from memory. Using the cognitive architecture ACT-R (Chapter 2) and connectionist constraint satisfaction models based on ECHO (Chapter 3), we implemented several computational models. The models shared the assumption that observed information can activate associated knowledge, but they differed with respect to how the sequentially made observations affected memory activation over time. The results of the models were compared to human data from two behavioral experiments in which we used a probe reaction task to track the availability of different explanations during a sequential diagnostic reasoning task. The basic results were consistent over both approaches: Comparing the probe reaction data to the models' results suggested that the availability of explanations in memory indeed varied as a function of the observed symptoms over time. Furthermore, the probe

reaction data was best fit by models in which the influence of observed information did not vary as a function of the number of observations (Chapter 2) and remained stable over time (Chapters 2 and 3).

Both modeling approaches increased the precision compared to mere verbal theories and supported our assumptions about memory activation. However, the approaches differed in the scope of their interpretability. The connectionist models showed how observed data could activate associated knowledge in a network. But how is such a network constructed and which aspects of memory does it reflect? Using ACT-R, and thereby adhering to the constraints set by the underlying memory theory, allowed for interpreting our results more functionally, in terms of general memory mechanisms. We concluded that observed symptoms that are currently in the focus of attention regulate the availability of associated explanations in long-term memory by spreading activation to these explanations. This component of memory activation reflects the usefulness of explanations in the current context.

In Chapter 4, we investigated how the influence of the current context interacts with a second factor that has been proposed to influence an item's availability in memory: its past usefulness (e.g., Anderson, 2007; Thomas, et al., 2008). We conducted a behavioral experiment in which both factors, past and present usefulness, were independently manipulated by a secondary task. Results of this experiment were compared to the predictions of an ACT-R model that was constructed based on our findings in Chapter 2. Participants' performance showed effects of both manipulations, as predicted by the model, suggesting that the past *and* the present usefulness determine the availability of diagnostic hypotheses in memory.

Why is it important to understand memory processes underlying diagnostic reasoning in so much detail? As, for example, discussed by Dougherty et al. (2010), many theories assume that "whatever takes place in the memory system is irrelevant to understanding judgment and decision-making behavior" (p. 337). Our results illustrate why such a simplifying assumption is short-sighted. By using a precise account of memory activation, which was based on general findings about memory mechanisms, we could show how taking into account the contribution of memory processes can lead to a better understanding of hypothesis generation. Such an understanding is not only interesting from a theoretical perspective, but it can, potentially, help to improve real-world decision making.

Consider, for example, the medical setting, where a doctor's ability to generate correct diagnoses can be of vital importance for a patient. The medical literature describes various pitfalls that frequently occur in this setting (for an overview see e.g., Klein, 2005). An understanding of the underlying cognitive mechanisms can help to develop training programs that might reduce such pitfalls. Take, for example, the representativeness heuristic (Tversky & Kahneman, 1974), where diagnosticians strongly rely on information available in the current context (e.g., a patient's symptoms) but seem to ignore information about the base-rate likelihood of the potential diagnoses. Our results suggest that this effect might be due to insufficient personal experience with the diagnoses' base rates. If real-world base rates are not represented in memory in terms of experienced frequencies, the past-experience component of memory

activation cannot correctly reflect the real-world base rate information. Consequently, memory activation can be dominated by the current-context component, resulting in behavior as described by the representativeness heuristic. Teaching base-rates in terms of natural frequencies rather than as abstract percentages (see e.g., Sedlmeier & Gigerenzer, 2001, for how that could be done) might help to reduce this pitfall, because it would allow for memory activation to adaptively provide the diagnosis that is most likely not only based on the current context but also based on past experience.

Decision Making Based on Information from Memory

Whereas in Chapters 2 to 4 we investigated how memory activation affects the availability of information in memory as a function of the past and present environment, in Chapter 5 we investigated how reasoners make decisions by exploiting the availability of memory contents. This investigation is directly related to the second dichotomy mentioned above: the question whether decisions can better be described by simple non-compensatory heuristics or by more complex compensatory decision making strategies. A non-compensatory heuristic has, for example, been proposed in terms of the recognition heuristic (Goldstein & Gigerenzer, 2002). This heuristic states that if only one of two alternatives is recognized, the reasoner will rely on recognition to infer the recognized alternative to have higher values on a given criterion, without using additional knowledge about the alternatives. In contrast to such a non-compensatory heuristic, compensatory decision models assume that people use additional knowledge about the alternatives (Glöckner & Betsch, 2008; Lee & Cummins, 2004). For example, when deciding which of two cities, one recognized and one not, is larger, the reasoner would decide for the recognized city according to the recognition heuristic. According to compensatory models of decision making, the reasoner would take into account additional knowledge about the cities' cues (e.g., does it have an airport?) and might, consequently, conclude that the recognized city is smaller.

As we discussed in Chapter 5, the comparison of these apparently opposing decision strategies has been proven difficult, because the strategies are described at varying levels of detail and are often underspecified relative to the empirical data against which they can be tested. ACT-R allowed us to tackle these issues by implementing several apparently opposing strategies within one modeling framework. This implementation required a high degree of precision and took into account the interplay of the decision strategies themself with, for example, perceptual, memory, and motor processes. Comparing the models to behavioral data from Pachur et al. (2008) showed that models that incorporated a combination of supposedly opposing strategies fit the data best. For example, even participants that always decided for the recognized alternative seemed to occasionally retrieve additional knowledge from memory.

The results of Chapter 5 illustrate again how cognitive architectures like ACT-R can be used to dissolve simplifying dichotomies and increase our understanding of detailed cognitive processes. Furthermore, the results highlight the importance of

taking into account the interaction of (deliberate) decision strategies with other aspects of cognition, like the availability of information in memory. For instance, in Pachur et al.'s experiments, all knowledge about the alternatives was taught in a very controlled setup, making it, for example, likely that the different cues about an alternative were equally available in memory. In a more natural setting, the availability of different pieces of information in memory will vary as a function of the environment, which can result in different outcomes for the same decision strategy, and might even cause the use of different decision strategies.

Conclusion

In this thesis I have shown how the precision provided by computational cognitive models can be used to better understand the cognitive processes underlying complex cognition. Our results illustrate why it is often not useful to construct and test binary oppositions, as it is done in many areas of research. We showed why apparently opposing aspects of cognition, like automatic and deliberate reasoning, or non-compensatory and compensatory decision making, should better be understood as complementary components. For me, the discovery of the respective contribution and interaction of these components represents the real challenge in understanding the human mind, and I hope that the work presented in this thesis represents a step towards solving this challenge. Have we, now in 2011, "homed in to the essential structure of the mind"? While undoubtedly progress has been made, I want to close with a quote from John Anderson: "we are still only a little ways into understanding the answer" (Anderson, 2007, p. 239).

Samenvatting

Stel je voor dat een vriend na een zonnige dag op het strand klaagt over hoofdpijn, een pijnlijke rode huid en jeuk. Hoe kun je uitvinden wat er met hem aan de hand is? Ga je een medisch naslagwerk raadplegen om een mogelijke diagnose voor zijn symptomen te vinden? Ga je over de fysieke processen in zijn lichaam nadenken die de oorzaak van de symptomen zouden kunnen zijn? Of ga je eerst de verschillende mogelijke oorzaken bedenken en deze tegen elkaar afwegen? Waarschijnlijk doe je niets van dit alles, maar verschijnt de diagnose 'verbrand' automatisch in je hoofd. Hoe is dit mogelijk?

Terwijl in het verleden veel onderzoek is gedaan naar bewuste denkprocessen en strategieën die mensen gebruiken om complexe problemen op te lossen, is er tegenwoordig veel interesse in onderzoek naar automatische denkprocessen. Zo publiceerden Dijksterhuis en collega's (2006) een artikel in het bekende tijdschrift Science, waarin ze lieten zien dat onbewust denken tot betere beslissingen kan leiden dan bewust nadenken. Maar, hoe werkt 'onbewust denken'? Welk mechanisme in ons brein maakt het mogelijk dat we aan 'verbrand' denken zodra we de bijbehorende symptomen zien? En hoe werkt de interactie tussen zulke automatische processen en bewuste denkprocessen? Een mogelijk antwoord op deze vragen vinden we in het functioneren van het menselijke geheugen: "the memory system […] makes most available those memories most likely to be needed" (Anderson, 2007, p. 109).

Het beginpunt van mijn promotieonderzoek was het idee dat automatische geheugenprocessen een belangrijk aspect van complexe cognitie vormen. Om precies te zijn was ik geïnteresseerd in de rol van automatische geheugenprocessen tijdens het stellen van diagnoses. Bij het stellen van een diagnose wordt er vanuit observaties geredeneerd naar mogelijk oorzaken. Een onderdeel hiervan is het genereren van hypotheses over wat deze mogelijke oorzaken zouden kunnen zijn. Het stellen van diagnoses is een belangrijk onderdeel van veel taken, zoals medische diagnoses, software debugging, wetenschappelijk onderzoek en sociale interacties. Toen ik aan dit proefschrift begon was mijn hypothese dat automatische geheugenprocessen diagnoses beschikbaar stellen die geassocieerd zijn met de huidige context. Deze informatie kan dan vervolgens gebruikt worden in bewuste redeneerprocessen. Hoewel dit idee op zich niet nieuw is, bestonden er nauwelijks precieze theorieën en experimenteel bewijs voor. Thomas en collega's stelden het zo in Psychological Review: "despite hypothesis generation's importance in understanding judgment, little empirical and even less theoretical work has been devoted to understanding the processes underlying hypothesis generation." (Thomas, et al., 2008, p. 174).

Gedragsexperimenten

Om het idee te testen dat informatie uit de omgeving geassocieerde informatie in ons geheugen activeert, hebben we eerst enkele gedragsexperimenten uitgevoerd. In deze experimenten hebben we de mate van associatie tussen medische symptomen en diagnoses in het geheugen gemanipuleerd. Om te kijken of deze manipulatie een

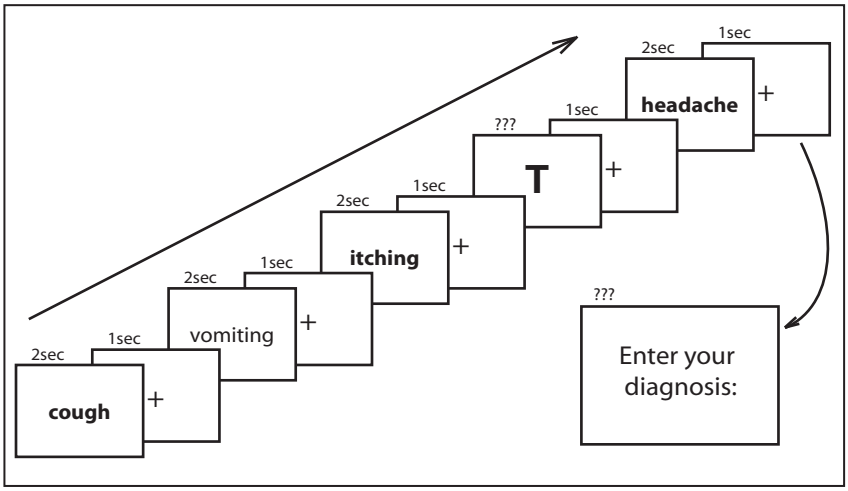
effect had, hebben we een zogenaamde 'probe reaction task' gebruikt. Dit ging als volgt: tijdens het presenteren van symptomen lieten we zo nu en dan een mogelijke diagnose zien. Als het inderdaad het geval is dat diagnoses in het geheugen geactiveerd worden door de symptomen, dan zouden proefpersonen sneller moeten reageren op bijvoorbeeld 'verbrand' (met symptomen zoals 'rode huid' en 'hoofdpijn') dan op 'zwanger', of 'huis'. Om een mogelijke invloed van bestaande kennis uit te sluiten hebben we in deze experimenten alleen gebruik gemaakt van kunstmatige medische kennis. Deze kennis bestond uit symptomen die veroorzaakt werden door mogelijke (ook zelfbedachte) chemicaliën. Deze chemicaliën werden beschreven door letters (bijvoorbeeld 'B' of 'W'), en deze letters gebruikten we dus ook in de probe reaction task (zie Figuur 7.1 voor een voorbeeld van de taak). Dit zorgde ervoor dat we geen verstorende effecten van bijvoorbeeld woordlengte of leessnelheid kregen.

In het algemeen ondersteunden de resultaten van deze experimenten onze theorie. Op diagnoses die overeenkwamen met alle tot nog toe getoonde symptomen werd bijvoorbeeld het snelste gereageerd, wat suggereert dat deze diagnoses beter beschikbaar waren in het geheugen dan diagnoses die niet overeenstemden met de symptomen. Ook werden reactietijden korter als er meer symptomen met een diagnose overeenkwamen. Dit suggereert dat de getoonde symptomen inderdaad de beschikbaarheid van hypothesen in het geheugen beïnvloeden. Bij een nadere inspectie van onze resultaten bleek echter dat er nog veel vragen onbeantwoord bleven. Het zou bijvoorbeeld zo kunnen zijn dat de reactietijdverschillen niet veroorzaakt werden door geheugenactivatie, maar alleen een bijproduct waren van bewuste redeneerprocessen. En, stel dat geheugenactivatie de resultaten veroorzaakte, hoe werkt dit dan precies?

Computationele Cognitieve Modellen

Het gebrek aan theoretische nauwkeurigheid die we bij de interpretatie van onze resultaten tegenkwamen is typerend voor verbale theorieën. Een mogelijke oplossing voor dit probleem is het gebruik van computationele cognitieve modellen. Dit soort modellen zijn computersimulaties van de processen die volgens een theorie gedrag veroorzaken (bijvoorbeeld van de geheugenprocessen die de activatie van hypothesen in het geheugen veroorzaken). Door de resultaten van een simulatie te vergelijken met gedragsdata van proefpersonen kunnen theorieën veel nauwkeuriger getest worden dan door alleen naar verbale voorspellingen te kijken. Wij hebben gebruik gemaakt van twee verschillende soorten cognitieve modellen: een connectionistisch model (ECHO) en een cognitieve architectuur (ACT-R). Ik zal het idee achter deze modellen hieronder kort uitleggen, voordat ik een samenvatting van de modelresultaten geef.

Een connectionistisch model is een netwerk van knopen, die door exciterende of inhiberende connecties met elkaar zijn verbonden. Onder andere Paul Thagard (2000) stelde voor dat dit soort netwerksimulaties gebruikt kunnen worden om integratie van informatie (bijvoorbeeld tijdens het stellen van diagnoses) te modelleren. In zijn computersimulatie ECHO representeert elke knoop in het netwerk één concept (bijvoorbeeld een geobserveerd symptoom of een mogelijke diagnose) en de



Figuur 7.1 Voorbeeld van een experiment met de 'probe reaction task' (zie Experiment 1 in Hoofdstuk 2). Tijdens het vertonen van symptomen wordt een probe gepresenteerd (T). De proefpersoon moet zo snel mogelijk aangeven of de probe een chemisch element is of niet. Probes variëren in de mate van associatie met de symptomen (T in dit voorbeeld is een diagnose die met alle getoonde symptomen overeen komt). De tijd die voor de reactie op de probe nodig is geeft aan hoe goed die diagnose in het geheugen beschikbaar is.

verbindingen tussen de knopen representeren de relaties tussen deze concepten. Als het netwerk wordt geïnitieerd, activeren of inhiberen de concepten elkaar. Uiteindelijk krijgt het concept dat het meest met alle andere concepten overeenkomt de meeste activatie. Als dit model toegepast wordt op het stellen van diagnoses kan het voorspellen hoe sterk een bepaalde diagnose door geobserveerde symptomen word geactiveerd.

Een andere manier van modellieren is het gebruik van cognitieve architecturen. Een cognitieve architectuur is eigenlijk een verzameling van theorieën en modellen die verschillende aspecten van cognitie weerspiegelen. Het doel is om beter te begrijpen hoe het brein het mogelijk maakt dat wij kunnen nadenken (Anderson, 2007). De onderzoeker kan gebruik maken van deze verzamelde kennis (bijvoorbeeld over hoe het menselijke geheugensysteem werkt) om zijn eigen specifieke vraag nader te onderzoeken (bijvoorbeeld, hoe beïnvloeden observaties de beschikbaarheid van diagnoses in het geheugen). In dit proefschrift heb ik gebruik gemaakt van de cognitieve architectuur die tegenwoordig het meest gebruikt wordt: ACT-R (Anderson, et al., 2004). ACT-R is bijzonder geschikt voor mijn onderzoek omdat het niet alleen een gedetailleerde theorie over het menselijke geheugen heeft, maar ook over de interactie tussen automatische geheugenprocessen en bewuste denkprocessen.

Geheugenactivatie en het Genereren van Hypothesen

In Hoofdstuk 2 en 3 hebben we onderzocht of en hoe informatie uit de omgeving geassocieerde informatie in ons geheugen activeert. Met behulp van ACT-R (Hoofdstuk 2) en ECHO (Hoofdstuk 3) hebben we verschillende cognitieve modellen geïmplementeerd. Alle modellen delen de aanname dat geobserveerde informatie geassocieerde informatie in ons geheugen activeert. De modellen verschillen echter in hoe deze activatie precies verloopt gedurende de tijd. De vergelijking van de modelresultaten met de gedragsdata van proefpersonen uit de hierboven uitgelegde 'probe reaction task' liet zien dat de beschikbaarheid van diagnoses in het geheugen inderdaad varieert als een functie van de geobserveerde informatie. Verder toonden de resultaten aan hoe observaties gedurende de tijd worden geïntegreerd.

Beide manieren van modelleren vergrootten de precisie vergeleken met verbale theorieën en ondersteunden onze aannamen over de rol van geheugenactivatie. De modellen verschillen echter in de mate waarin de resultaten geïnterpreteerd kunnen worden. De connectionistische modellen laten zien hoe geobserveerde symptomen geassocieerde diagnoses in een netwerk kunnen activeren. Maar welke aspecten van het geheugen worden door het netwerk gereflecteerd? ACT-R maakt het mogelijk om de resultaten binnen een gedetailleerde geheugentheorie te interpreteren. Het blijkt dat informatie waaraan we aandacht besteden (in dit geval observaties in ons werkgeheugen) automatisch geassocieerde informatie (zoals diagnoses) in het langetermijngeheugen activeren. Op deze manier beïnvloedt de huidige context de beschikbaarheid van hypothesen in het geheugen.

Geheugentheorieën tonen aan dat de beschikbaarheid van informatie in ons geheugen niet alleen afhankelijk is van de huidige context, maar ook van onze ervaring met deze informatie (b.v. Anderson, 2007; Thomas, et al., 2008). Hoe vaker je de informatie al uit je geheugen hebt gehaald, en hoe recenter dat was, des te sterker is de informatie in je geheugen aanwezig, en des te makkelijker kan deze geactiveerd worden. In Hoofdstuk 4 hebben we onderzocht of ook bij het genereren van hypothesen beide factoren, ervaring en context, een rol spelen. In een gedragsexperiment moesten proefpersonen diagnoses voor medische symptomen stellen. Tijdens het genereren van deze diagnoses, moest een tweede taak worden uitgevoerd, met behulp waarvan we beide factoren onafhankelijk van elkaar konden manipuleren. Gedragsdata in de verschillende condities hebben we vergeleken met de voorspellingen van een ACT-R model. Zoals het model voorspeld had, werd de prestatie op de diagnosetaak beïnvloed door zowel ervaring als context. Deze resultaten ondersteunen dus het idee dat zowel ervaring als de huidige context de beschikbaarheid van diagnoses in het geheugen beïnvloeden en laten zien hoe geheugenmechanismes deze invloed veroorzaken.

Je kunt je afvragen waarom het zo belangrijk is om geheugenmechanismes zo goed te begrijpen. Zoals Dougherty et al. (2010) bijvoorbeeld beschrijft, veronderstellen theorieën vaak dat "whatever takes place in the memory system is irrelevant to understanding judgment and decision-making behavior" (p. 337). Onze resultaten laten zien waarom deze aanname kortzichtig is. Geheugenprocessen beïnvloeden welke informatie we ons, afhankelijk van onze ervaring en de huidige context,

kunnen herinneren. Daarmee beïnvloeden ze ook welke informatie we voor bewuste denkprocessen beschikbaar hebben. Door beter te begrijpen hoe geheugenprocessen de beschikbaarheid van informatie reguleren kunnen we het stellen van diagnoses beter begrijpen. Dit is niet alleen interessant vanuit een theoretisch perspectief, maar het kan wellicht ook helpen om het nemen van beslissingen in de praktijk te verbeteren.

Denk bijvoorbeeld aan de medische diagnose, waar correcte beslissingen van levensbelang voor de patiënt kunnen zijn. De medische literatuur beschrijft een aantal typische fouten bij het stellen van diagnoses (zie b.v. Klein, 2005). Als we de cognitieve mechanismen tijdens het diagnosticeren beter begrijpen, kunnen we meer over de oorzaken van deze fouten leren. Deze kennis kan gebruikt worden om trainingsprogramma's te ontwikkelen die helpen de fouten te voorkomen. Een typische fout is bijvoorbeeld de 'representativeness heuristic', waarbij een diagnose te sterk wordt beïnvloed door de huidige context, zoals de symptomen van een patiënt), en er te weinig rekening wordt gehouden met hoe vaak een diagnose in het algemeen voorkomt. Onze resultaten doen vermoeden dat deze fout wordt veroorzaakt door te weinig persoonlijke ervaring met een diagnose. Als de persoonlijke ervaring van een arts een incorrecte afspiegeling is van hoe vaak een diagnose in de werkelijkheid voorkomt, dan kan de geheugenactivatie dit aspect niet gebruiken en wordt de huidige context belangrijker. Deze fout kan wellicht voorkomen worden door het aantal optredens van een diagnose niet alleen in de vorm van abstracte getallen te trainen, maar ook in de vorm van natuurlijke frequenties.

Beschikbaarheid van Informatie en het Nemen van Beslissingen

In de Hoofdstukken 2, 3 en 4 hebben we geheugenprocessen onderzocht die de beschikbaarheid van informatie beïnvloeden. In Hoofdstuk 5 zijn we een stap verder gegaan en hebben we gekeken hoe de beschikbaarheid van informatie samenhangt met het nemen van beslissingen. Op dit moment is er veel discussie of het nemen van beslissingen beter kan worden begrepen in termen van eenvoudige heuristieken of als een complexer proces waarbij verschillende feiten tegen elkaar worden afgewogen.

Een bekende heuristiek is de 'recognition heuristic' (Goldstein & Gigerenzer, 2002). Deze heuristiek kan bijvoorbeeld gebruikt worden om de vraag te beantwoorden welke van twee alternatieven (twee steden zoals 'Leipzig' en 'Chemnitz') hogere waarden op een bepaald criterium (inwoners) heeft. De heuristiek veronderstelt dat als we één van de alternatieven herkennen (Leipzig) en de andere niet (Chemnitz), we voor het bekende alternatief kiezen zonder er verder over na te denken. Daartegenover staan theorieën die veronderstellen dat bij een beslissing meer informatie wordt gebruikt. Zo zou je kunnen bedenken dat, voor zover jij weet, Leipzig geen internationaal vliegveld en geen bekende voetbalploeg heeft. Omdat deze feiten tegen het idee ingaan dat Leipzig een grote stad is zou je misschien voor Chemnitz kiezen, terwijl je helemaal niets over Chemnitz weet.

Zoals een groot aantal publicaties laat zien is de vergelijking van deze schijnbaar tegengestelde strategieën vaak moeilijk. Dit komt omdat strategieën in de literatuur op een verschillende niveau beschreven worden en vaak alleen verbale voorspellingen maken. In Hoofdstuk 5 beschrijven we hoe ACT-R helpt deze problemen aan te pakken. Het gebruik van een cognitieve architectuur maakt het namelijk mogelijk om verschillende strategieën binnen één theoretisch kader te implementeren. Deze implementatie leidt niet alleen tot precieze voorspellingen, maar houdt ook rekening met de interactie tussen de strategieën en andere aspecten van cognitie, zoals geheugenprocessen. De vergelijking van de voorspellingen met gedragsdata van twee gepubliceerde experimenten (Pachur, et al., 2008) liet zien dat combinaties van schijnbaar tegengestelde strategieën het gedrag het beste voorspellen: Bijvoorbeeld, zelfs proefpersonen die uiteindelijk altijd voor de bekende alternatieven kozen, bleken gedeeltelijk informatie over deze alternatieven uit het geheugen te halen.

Conclusie

Al in de jaren 70 waarschuwde Allen Newell ervoor dat verbale theorieën en het vergelijken van steeds twee alternatieve aannamen niet voldoende is om ons begrip van cognitie te vergroten (A. Newell, 1973). Desondanks zijn verbale theorieën en binaire aannamen nog steeds heel populair binnen de wetenschap. In dit proefschrift heb ik laten zien hoe de precisie van cognitieve modellen gebruikt kan worden om cognitieve processen beter te begrijpen. Onze resultaten maken duidelijk dat schijnbaar tegenstrijdige aspecten van cognitie, zoals automatische en bewuste denkprocessen, of het gebruik van eenvoudige heuristieken en complexere strategieën bij het nemen van beslissingen, vaak complementair zijn. We hebben bijvoorbeeld aangetoond dat automatische activeringsprocessen de beschikbaarheid van informatie in het geheugen reguleren en daarmee beïnvloeden welke informatie door bewuste denkprocessen gebruikt kan worden. Het beter begrijpen van dit soort interacties is één van de grote uitdagingen in het begrijpen van menselijke cognitie. Ik hoop dat dit proefschrift een goede stap in deze richting is.

Acknowledgements

The story of this thesis could not have been turned into a success without the help and encouragement of numerous people.

First of all, I want to thank Martin Bauman. Martin, you gave me the initial idea for this thesis. Thanks for that and for sharing your scientific knowledge, your enthusiasm, and countless valuable tips.

Second, I want to thank Josef Krems. Without you I could not have started working on this thesis. You supervised the early stages, coauthored several papers, and made it possible for me to attend numerous conferences and of course the ACT-R summer school. Finally, you were a member of the reading committee. Thank you for all that.

Frank Ritter, you had a strong impact on my thesis early on. You taught me a lot about secret weapons, scientific writing, and cognitive modeling. Thanks for that, and for telling me about Newell.

Georg Jahn, you helped me to actually get started with modeling. Thanks for taking me on the route to connectionism with you and for the countless interesting discussions about abductive reasoning.

While ACT-R seemed to be the most promising modeling framework from the beginning, I might not have started to use it, had it not been for our ACT-R self-help group. I want to thank everyone who was involved in this, but especially you, Matthias, for bringing it into life and you, Udo, for keeping it alive by always coming up with the correct solutions and sharing them with the rest of us.

Following the trace of ACT-R brings me to the 2008 ACT-R summer school. I want to thank my mentor Ion Juvina, all the instructors, and my co-students. Attending the summer school and the subsequent workshop was life changing for me in more ways than I would ever have expected.

One of the results of the summer school was that I could spend 3.5 months at CMU to work with Niels Taatgen and Christian Lebiere. Thanks Niels and Christian for taking all this time and teaching me so much.

Of course, Niels, I want to thank you for so many other things. Thanks for taking over the supervision of my project at an advanced stage and giving me a place in Groningen. Thanks for all those useful modeling tips, for helping me to get my papers written up, and to cope with the reviewers' comments. I do not know how I could have ever finished this thesis without your help. And thanks to you and Steffi, for all those entertaining game nights.

Another result of using ACT-R was that I started a project with you, Julian Marewski. What first seemed like a side project became a major part of my dissertation. Thanks for our successful cooperation.

Fokie Cnossen, thanks for your help, interest, and enthusiasm. Who knows what would have become of our project, if you had not insisted on trying out that one additional variation of the experiment?

Last but not least in this list, I want to thank the other two members of my committee: Rineke Verbrugge and Coty Gonzalez. Thanks for reading and evaluating my thesis and thanks for taking the (long) trip to my defense in Groningen.

While the people mentioned above where directly involved with my thesis, I want to thank a number of other people who helped me in many different ways during my PhD time.

Hedderik, thank you for so many things. To mention some: A job, an office, and always-useful comments on my thesis and on whatever else we talked about.

Jacolien, thanks for being the best office mate one could wish for. And the best cat sitter, of course.

Thanks to you, Simone and Steffi. For keeping my German fluent, for all those wonderful Stammtischmeetings, and for sharing the secrets that one needs to know when being married to a Dutch scientist.

And Jacolien and Simone, thanks for being willing to defend me as my 'paranimfen'.

The cognitive modeling group, thanks for your useful comments and for making me believe I can do it.

Thanks also to all my colleagues and students in Chemnitz and Groningen who where not mentioned yet. You provided so much useful input to this thesis and you always created a working environment in which it was fun to work, despite the challenges that come with doing a PhD.

I cannot close, before thanking my family and friends; two- and four-legged. You always encouraged me, you took my mind away from work whenever I needed it, and you gave me the energy to keep going.

And you, Jelmer. Where should I start thanking you? Thank you for everything! Without you, in every possible way, I wouldn't be where I am now.

References

Anderson, J. R. (1976). *Language, Memory, and Thought*: Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1990). *The Adaptive Character of Thought*: Hillsdale, NJ: Erlbaum.

Anderson, J. R. (2007). *How Can the Human Mind Occur in the Physical Universe?* New York, NY: Oxford University Press.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review, 111*, 1036-1060.

Anderson, J. R., Bothell, D., Lebiere, C., & Matessa, M. (1998). An integrated theory of list memory. *Journal of Memory and Language, 38*, 341-380.

Anderson, J. R., Fincham, J. M., Qin, Y., & Stocco, A. (2008). A central circuit of the mind. *Trends in Cognitive Sciences, 12*, 136-143.

Anderson, J. R., & Lebiere, C. (1998). The atomic components of thought. Mahwah: NJ: Erlbaum.

Anderson, J. R., & Schooler, L. (1991). Reflections of the environment in memory. *Psychological Science, 2*, 396-408.

Arocha, J. F., & Patel, V. L. (1995). Construction-integration theory and clinical reasoning. In C. Weaver, S. Mannes & C. Fletcher (Eds.), *Discourse Comprehension: Essays in Honor of Walter Kintsch* (pp. 359-382). Hillsdale, NJ: Erlbaum.

Arocha, J. F., Wang, D., & Patel, V. L. (2005). Identifying reasoning strategies in medical decision making: A methodological guide. *Journal of Biomedical Informatics, 38*, 154–171.

Barrows, H., Norman, G., Neufeld, V., & Feightner, J. (1982). The clinical reasoning of randomly selected physicians in general medical practice. *Clinical and Investigative Medicine, 5*, 49-55.

Baumann, M. R. K. (2001). *Die Funktion des Arbeitsgedächtnisses beim abduktiven Schließen: Experimente zur Verfügbarkeit der mentalen Repräsentation erklärter und nicht erklärter Beobachtungen [The function of working memory in abductive reasoning: Experiments on the availability of the mental representation of explained and unexplained observations].* Doctoral dissertation. Chemnitz University of Technology, Chemnitz. Available from http://archiv.tu-chemnitz.de/pub/2001/0071.

Baumann, M. R. K., Krems, J. F., & Ritter, F. E. (2010). Learning from examples does not prevent order effects in belief revision. *Thinking & Reasoning, 16*, 98-130.

Baumann, M. R. K., Mehlhorn, K., & Bocklisch, F. (2007). The activation of hypotheses during abductive reasoning. In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Cognitive Science Society* (pp. 803-808). Austin, TX: Cognitive Science Society.

Bergert, F. B., & Nosofsky, R. M. (2007). A response-time approach to comparing generalized rational and take-the-best models of decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 107-129.

Berman, M., Jonides, J., & Lewis, R. (2009). In search of decay in verbal short-term memory. *Learning, Memory, 35*, 317-333.

Böhm, U., & Mehlhorn, K. (2009). The influence of spreading activation on memory retrieval in sequential diagnostic reasoning. In A. Howes, D. Peebles & R. Cooper

(Eds.), *Proceedings of the 9th International Conference on Cognitive Modeling*. Manchester, UK.

Borst, J. P., Taatgen, N. A., & van Rijn, H. (2010). The problem state: A cognitive bottleneck in multitasking. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 36*, 363-382.

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In R. L. Launer & G. N. Wilkinson (Eds.), *Robustness in Statistics*. New York, NY: Academic Press.

Brandstätter, E., Gigerenzer, G., & Hertwig, R. (2006). The priority heuristic: Making choices without trade-offs. *Psychological Review, 113*, 409-432.

Bröder, A. (2003). Decision making with the "adaptive toolbox": Influence of environmental structure, intelligence, and working memory load. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 611-625.

Bröder, A., & Eichler, A. (2006). The use of recognition information and additional cues in inferences from memory. *Acta Psychologica, 121*, 275-284.

Bröder, A., & Gaissmaier, W. (2007). Sequential processing of cues in memory-based multiattribute decisions. *Psychonomic Bulletin & Review, 14*, 895-900.

Bröder, A., & Schiffer, S. (2003). Take The Best versus simultaneous feature matching: Probabilistic inferences from memory and effects of reprensentation format. *Journal of Experimental Psychology: General, 132*, 277-293.

Bröder, A., & Schiffer, S. (2006). Stimulus format and working memory in fast and frugal strategy selection. *Journal of Behavioral Decision Making, 19*, 361-380.

Bruner, J. S. (1957). On perceptual readiness. *Psychological Review, 64*, 123-152.

Busemeyer, J. R., & Myung, I. J. (1992). An adaptive approach to human decision making: Learning theory, decision theory, and human performance. *Journal of Experimental Psychology: General, 121*, 177-184.

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review, 100*, 432-459.

Busemeyer, J. R., & Wang, Y. M. (2000). Model comparisons and model selections based on generalization criterion methodology. *Journal of Mathematical Psychology, 44*, 171-189.

Bylander, T., Allemang, D., Tanner, M., & Josephson, J. (1991). The computational complexity of abduction. *Artificial Intelligence, 49*, 25-60.

Byrne, M. D., & Anderson, J. R. (2001). Serial modules in parallel: The psychological refractory period and perfect time-sharing. *Psychological Review, 108*, 847–869.

Chechile, R. A. (2003). Mathematical tools for hazard function analysis. *Journal of Mathematical Psychology, 47*, 478-494.

Chinn, C., & Brewer, W. (1998). An empirical test of a taxonomy of responses to anomalous data in science. *Journal of Research in Science Teaching, 35*, 623-654.

Chuderski, A., Stettner, Z., & Orzechowski, J. (2006). Modeling individual differences in a working memory search task. In D. Fum, F. Del Missier & A. Stocco (Eds.), *Proceedings of the 7th International Conference on Cognitive Modeling* (pp. 74-79). Trieste, Italy.

Cokely, E. T., & Kelley, C. M. (2009). Cognitive abilities and superior decision making under risk: A protocol analysis and process model evaluation. *Judgment and Decision Making, 4*, 20-33.

Cokely, E. T., Parpart, P., & Schooler, L. J. (2009). On the link between cognitive control and heuristic processes. In N. A. Taatgen & H. Van Rijn (Eds.), *Proceedings of the 31th Annual Conference of the Cognitive Science Society* (pp. 2926–2931). Austin, TX: Cognitive Science Society.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*, 87-185.

Davis-Stober, C. P., Dana, J., & Budescu, D. V. (2010). Why recognition is rational: Optimality results on single-variable decision rules. *Judgment and Decision Making, 5*, 216-229.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34*, 571-582.

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81*, 95-106.

De Neys, W. (2006). Dual processing in reasoning. *Psychological Science, 17*, 428-433.

Dijksterhuis, A., Bos, M. W., Nordgren, L. F., & Van Baaren, R. B. (2006). On making the right choice: The deliberation-without-attention effect. *Science, 311*, 1005.

Dijksterhuis, A., & Nordgren, L. F. (2006). A theory of unconscious thought. *Perspectives on Psychological Science, 1*, 95-109.

Dougherty, M. R. P., Franco-Watkins, A. M., & Thomas, R. (2008). Psychological plausibility of the theory of probabilistic mental models and the fast and frugal heuristics. *Psychological Review, 115*, 199-213.

Dougherty, M. R. P., Gettys, C. F., & Ogden, E. E. (1999). MINERVA-DM: A memory processes model for judgments of likelihood. *Psychological Review, 106*, 180-209.

Dougherty, M. R. P., Gettys, C. F., & Thomas, R. P. (1997). The role of mental simulation in judgments of likelihood. *Organizational Behavior and Human Decision Processes, 70*, 135-148.

Dougherty, M. R. P., & Hunter, J. E. (2003a). Hypothesis generation, probability judgment, and individual differences in working memory capacity. *Acta Psychologica, 113*, 263-282.

Dougherty, M. R. P., & Hunter, J. E. (2003b). Probability judgment and subadditivity: The role of working memory capacity and constraining retrieval. *Memory & Cognition, 31*, 968-982.

Dougherty, M. R. P., & Sprenger, A. M. (2006). The influence of improper sets of information on judgment: How irrelevant information can bias judged probability. *Journal of Experimental Psychology: General, 135*, 262-281.

Dougherty, M. R. P., Thomas, R., & Lange, N. (2010). Toward an integrative theory of hypothesis generation, probability judgment, and hypothesis testing. *Psychology of Learning and Motivation, 52*, 299-342.

Drewitz, U., & Thüring, M. (2009). Modeling the confidence of predictions: A time based approach. In A. Howes, D. Peebles & R. Cooper (Eds.), *Proceedings of the 9th*

International Conference of Cognitive Modeling. Manchester, UK.
Einhorn, H. J., & Hogarth, R. M. (1975). Unit weighting schemes for decision making. *Organizational Behavior and Human Performance, 13*, 171-192.
Einhorn, H. J., Kleinmuntz, D. N., & Kleinmuntz, B. (1979). Linear regression and process-tracing models of judgment. *Psychological Review, 86*, 465-485.
Elstein, A. S., Shulman, L. S., & Sprafka, S. A. (1978). *Medical Problem Solving: An Analysis of Clinical Reasoning*. Cambridge, MA: Harvard University Press.
Erdfelder, E., Küpper-Tetzel, C. E., & Mattern, S. D. (2011). Threshold models of recognition and the recognition heuristic. *Judgment and Decision Making, 6*, 7-22.
Ericsson, K., & Kintsch, W. (1995). Long-term working memory. *Psychological Review, 102*, 211-244.
Evans, J. S. B. T. (2006). The heuristic-analytic theory of reasoning: Extension and evaluation. *Psychonomic Bulletin and Review, 13*, 378–395.
Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology, 59*, 255-278.
Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition, 28*, 3-71.
Ford, J. K., Schmitt, N., Schechtman, S. L., Hults, B. M., & Doherty, M. L. (1989). Process tracing methods: Contributions, problems, and neglected research questions. *Organizational Behavior and Human Decision Processes, 43*, 75-117.
Fum, D., Del Missier, F., & Stocco, A. (2007). The cognitive modeling of human behavior: Why a model is (sometimes) better than 10,000 words. *Cognitive Systems Research, 8*, 135–142.
Gaissmaier, W., & Marewski, J. N. (2011). Forecasting elections with mere recognition from small, lousy samples: A comparison of collective recognition, wisdom of crowds, and representative polls. *Judgment and Decision Making, 6*, 73-88.
Gaissmaier, W., Schooler, L. J., & Mata, R. (2008). An ecological perspective to cognitive limits: Modeling environment-mind interactions with ACT-R. *Judgment and Decision Making, 3*, 278-291.
Gaissmaier, W., Schooler, L. J., & Rieskamp, J. (2006). Simple predictions fueled by capacity limitations: When are they successful? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*, 966-982.
Gettys, C., Pliske, R., Manning, C., & Casey, J. (1987). An evaluation of human act generation performance. *Organizational Behavior and Human Decision Processes, 39*, 23-51.
Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky. *Psychological Review, 103*, 592-596.
Gigerenzer, G. (1998). Surrogates for theories. *Theory and Psychology, 8*, 195-204.
Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science, 1*, 107-143.
Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103*, 650-669.
Gigerenzer, G., & Goldstein, D. G. (2011). The recognition heuristic: A decade of research. *Judgment and Decision Making, 6*, 100-121.

Gigerenzer, G., Hoffrage, U., & Goldstein, D. G. (2008). Fast and frugal heuristics are plausible models of cognition: Reply to Dougherty, Franco-Watkins, and Thomas (2008). *Psychological Review, 115*, 230.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98*, 506-528.

Glöckner, A., & Betsch, T. (2008). Modeling option and strategy choices with connectionist networks: Towards an integrative model of automatic and deliberate decision making. *Judgement and Decision Making, 3*, 215-228.

Glöckner, A., & Bröder, A. (2011). Processing of recognition information and additional cues: A model-based analysis of choice, confidence, and response time. *Judgment and Decision Making, 6*, 23-42.

Glöckner, A., & Hodges, S. D. (2011). Parallel constraint satisfaction in memory-based decisions. *Experimental Psychology (formerly Zeitschrift für Experimentelle Psychologie), 58*, 180-195.

Gluck, K. A. (2010). Cognitive architectures for human factors in aviation. In E. Salas & D. Maurino (Eds.), *Human Factors in Aviation, 2nd Edition* (pp. 375-400). New York, NY: Elsevier.

Gluck, K. A., Ball, J. T., & Krusmark, M. A. (2007). Cognitive control in a computational model of the Predator pilot. In W. D. Gray (Ed.), *Integrated Models of Cognitive Systems* (pp. 13-28). New York, NY: Oxford University Press.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience, 30*, 535-574.

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review, , 109*, 75-90.

Goldstein, D. G., & Gigerenzer, G. (2011). The beauty of simple models: Themes in recognition heuristic research. *Judgment and Decision Making, 6*, 392-395.

Grainger, J., & Jacobs, A. M. (1996). Orthographic processing in visual word recognition: A multiple read-out model. *Psychological Review, 103*, 518-565.

Gronlund, S. D., & Ratcliff, R. (1989). Time course of item and associative information: Implications for global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 846-858.

Hagmeyer, Y., & Waldmann, M. R. (2002). A constraint satisfaction model of causal learning and reasoning. In W. D. Gray & C. D. Schunn (Eds.), *Proceedings of the 24th Annual Conference of the Cognitive Science Society* (pp. 405-410). Mahwah, NJ: Erlbaum.

Hauser, J. R., & Wernerfelt, B. (1990). An evaluation cost model of consideration sets. *Journal of Consumer Research, 16*, 393-408.

Hertwig, R., Herzog, S., Schooler, L., & Reimer, T. (2008). Fluency heuristic: A model of how the mind exploits a by-product of information retrieval. *Learning, Memory, 34*, 1191-1206.

Higgins, E. T. (1996). Knowledge activation: Accessibility, applicability, and salience. In E. T. Higgins & A. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 133-168). New York, NY: Guilford Press.

Hilbig, B. E. (2008). Individual differences in fast-and-frugal decision making:

Neuroticism and the recognition heuristic. *Journal of Research in Personality, 42*, 1641-1645.

Hilbig, B. E., Erdfelder, E., & Pohl, R. F. (2010). One-reason decision making unveiled: A measurement model of the recognition heuristic. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 123-134.

Hilbig, B. E., & Pohl, R. F. (2009). Ignorance- versus evidence-based decision making: A decision time analysis of the recognition heuristic. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 1296-1305.

Hintzman, D. L., & Curran, T. (1994). Retrieval dynamics of recognition and frequency judgments: Evidence for separate processes of familiarity and recall. *Journal of Memory and Language, 33*, 1-18.

Hochman, G., Ayal, S., & Glöckner, A. (2010). Physiological arousal in processing recognition information: Ignoring or integrating cognitive cues. *Judgment and Decision Making, 5*, 285-299.

Hoffrage, U. (2011). Recognition judgments and the performance of the recognition heuristic depend on the size of the reference class. *Judgment and Decision Making, 6*, 43-57.

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology, 24*, 1-55.

Huber, O. (1989). Information-processing operators in decision making. In H. Montgomery & O. Svenson (Eds.), *Process and Structure in Human Decision Making* (pp. 3-21). New York, NY: Wiley.

Jacobs, A. M., & Grainger, J. (1994). Models of visual word recognition: Sampling the state of the art. *Journal of Experimental Psychology: Human Perception and Performance, 20*, 1311-1334.

Jacoby, L. L., & Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General, 110*, 306-340.

Johnson, E. J., Schulte-Mecklenbeck, M., & Willemsen, M. C. (2008). Process models deserve process data: Comment on Brandstätter, Gigerenzer, and Hertwig (2006). *Psychological Review, 115*, 263-272.

Johnson, T. R., & Krems, J. F. (2001). Use of current explanations in multicausal abductive reasoning. *Cognitive Science, 25*, 903-939.

Jonides, J., Lewis, R. L., Nee, D. E., Lustig, C. A., Berman, M. G., & Moore, K. S. (2008). The mind and brain of short-term memory. *Annual Review of Psychology, 59*, 193-224.

Joseph, G., & Patel, V. (1990). Domain knowledge and hypothesis genenation in diagnostic reasoning. *Medical Decision Making, 10*, 31-44.

Josephson, J. R., & Josephson, S. G. (1996). *Abductive Inference: Computation, Philosophy, Technology*. New York, NY: Cambridge University Press.

Just, M., & Carpenter, P. (1987). *The Psychology of Reading and Language Comprehension.* Boston, MA: Allyn and Bacon.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist, 58*, 697-720.

Kahneman, D., & Tversky, A. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124-1131.

Keren, G., & Schul, Y. (2009). Two is not always better than one: A critical evaluation of two-system theories. *Perspectives in Psychological Science, 4*, 533–550.

Kim, N., & Keil, F. (2003). From symptoms to causes: Diversity effects in diagnostic reasoning. *Memory & Cognition, 31*, 155-165.

Kintsch, W. (1998). *Comprehension: A Paradigm for Cognition*. New York, NY: Cambridge University Press.

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science, 12*, 1-48.

Klein, J. G. (2005). Five pitfalls in decisions about diagnosis and prescribing. *British Medical Journal, 330*, 781-783.

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review, 100*, 609-639.

Lebiere, C., & Anderson, J. R. (1993). A connectionist implementation of the ACT-R production system *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society* (pp. 635-640). New York, NY: Erlbaum.

Lee, M., & Cummins, T. (2004). Evidence accumulation in decision making: Unifying the 'take the best' and the 'rational' models. *Psychonomic Bulletin & Review, 11*, 343-352.

Lewandowsky, S. (1993). The rewards and hazards of computer simulations. *Psychological Science, 4*, 236-243.

Lewandowsky, S., Oberauer, K., & Brown, G. D. A. (2009). No temporal decay in verbal short-term memory. *Trends in Cognitive Sciences, 13*, 120-126.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review, 95*, 492-527.

Lovett, M. C., & Anderson, J. R. (1996). History of success and current context in problem solving. *Cognitive Psychology, 31*, 168-217.

Lovett, M. C., Daily, L. Z., & Reder, L. M. (2000). A source activation theory of working memory: Cross-task prediction of performance in ACT-R. *Cognitive Systems Research, 1*, 99-118.

Lyon, D. R., Gunzelmann, G., & Gluck, K. A. (2008). A computational model of spatial visualization capacity. *Cognitive Psychology, 57*, 122-152.

Marewski, J. N. (2008). *Ecologically Rational Strategy Selection.* Doctoral Dissertation, Free University, Berlin, Germany.

Marewski, J. N. (2010). On the theoretical precision and strategy selection problem of a single strategy approach: A comment on Glöckner, Betsch, and Schindler (2010). *Journal of Behavioral Decision Making, 23*, 463-467.

Marewski, J. N., Gaissmaier, W., & Gigerenzer, G. (2010a). Good judgments do not require complex cognition. *Cognitive Processing, 11*, 103-121.

Marewski, J. N., Gaissmaier, W., & Gigerenzer, G. (2010b). We favor formal models of heuristics rather than lists of loose dichotomies: A reply to Evans and Over. *Cognitive processing, 11*, 177-179.

Marewski, J. N., Gaissmaier, W., Schooler, L. J., Goldstein, D. G., & Gigerenzer, G.

(2009). Do voters use episodic knowledge to rely on recognition? In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 2232–2237). Austin, TX: Cognitive Science Society.

Marewski, J. N., Gaissmaier, W., Schooler, L. J., Goldstein, D. G., & Gigerenzer, G. (2010). From recognition to decisions: Extending and testing recognition-based models for multialternative inference. *Psychonomic Bulletin & Review, 17*, 287-309.

Marewski, J. N. & Mehlhorn, K. (2011). Using the ACT-R architecture to specify 39 quantitative process models of decision making. *Judgment and Decision Making, 6*, 439-519.

Marewski, J. N., & Olsson, H. (2009). Beyond the null ritual: Formal modeling of psychological processes. *Journal of Psychology, 217*, 49–60.

Marewski, J. N., Pohl, R. F., & Vitouch, O. (2010). Special Issue: Recognition processes in inferential decision making. *Judgement and Decision Making, 5*.

Marewski, J. N., Pohl, R. F., & Vitouch, O. (2011a). Special Issue: Recognition processes in inferential decision making (II). *Judgement and Decision Making, 6*.

Marewski, J. N., Pohl, R. F., & Vitouch, O. (2011b). Special Issue: Recognition processes in inferential decision making (III). *Judgement and Decision Making, 6*.

Marewski, J. N., & Schooler, L. J. (2011). Cognitive niches: An ecological model of strategy selection. *Psychological Review, 118*, 393-437.

Marewski, J. N., Schooler, L. J., & Gigerenzer, G. (2010). Five principles for studying people's use of heuristics. *Acta Psychologica Sinica, 42*, 72-87.

Mata, R., Schooler, L. J., & Rieskamp, J. (2007). The aging decision maker: Cognitive aging and the adaptive selection of decision strategies. *Psychology and Aging, 22*, 796-810.

McCloy, R., Beaman, C. P., & Smith, P. T. (2008). The Relative Success of Recognition Based Inference in Multichoice Decisions. *Cognitive Science, 32*, 1037-1048.

McElree, B., Dolan, P. O., & Jacoby, L. L. (1999). Isolating the contributions of familiarity and source information to item recognition: A time course analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 563-582.

Mehle, T. (1982). Hypothesis generation in an automobile malfunction inference task. *Acta Psychologica, 52*, 87-106.

Mehlhorn, K., Baumann, M. R. K., & Bocklisch, F. (2008). Activation or inhibition? Why reasoners are not blind for alternative explanations. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Cognitive Science Society*, (pp. 2083-2088). Austin, TX: Cognitive Science Society.

Mehlhorn, K., & Jahn, G. (2009). Modeling sequential information integration with parallel constraint satisfaction. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 2469-2474). Austin, TX: Cognitive Science Society.

Mehlhorn, K., Taatgen, N. A., Lebiere, C., & Krems, J. F. (2011). Memory activation and the availability of explanations in sequential diagnostic reasoning. *Journal of Experimantal Psychology: Learning, Memory, and Cognition, 37*, 1391-1411.

Meyer, D. E., & Kieras, D. E. (1997). A computational theory of executive cognitive processes and multiple-task performance: Part 1. Basic mechanisms. *Psychological*

Review, 104, 3-65.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology, 90*, 227-234.

Nellen, S. (2003). The use of the "take-the-best" heuristic under different conditions, modeled with ACT-R. In F. Detje, D. Dörner & H. Schaub (Eds.), *Proceedings of the Fifth International Conference on Cognitive Modeling* (pp. 171–176). Bamberg: Universitätsverlag Bamberg.

Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In W. G. Chase (Ed.), *Visual Information Processing* (pp. 283-308). San Diego, CA: Academic Press.

Newell, A. (1990). *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

Newell, A. (1992). SOAR as a unified theory of cognition: Issues and explanations. *Behavioral and Brain Sciences, 15*, 464-492.

Newell, B. R. (2005). Re-visions of rationality? *Trends in Cognitive Sciences, 9*, 11-15.

Newell, B. R., & Fernandez, D. (2006). On the binary quality of recognition and the inconsequentiality of further knowledge: Two critical tests of the recognition heuristic. *Journal of Behavioral Decision Making, 19*, 333-346.

Newell, B. R., & Lee, M. D. (in press). The right tool for the job? Comparing an evidence accumulation and a naïve strategy selection model of decision making. *Journal of Behavioral Decision Making*.

Newell, B. R., & Shanks, D. R. (2004). On the role of recognition in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 923-935.

Newell, B. R., Weston, N. J., & Shanks, D. R. (2003). Empirical tests of a fast-and-frugal heuristic: Not everyone "takes-the-best". *Organizational Behavior and Human Decision Processes, 91*, 82-96.

Nosofsky, R. M., & Bergert, F. B. (2007). Limitations of exemplar models of multi-attribute probabilistic inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 999-1019.

Oberauer, K. (2002). Access to information in working memory: Exploring the focus of attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*, 411-421.

Oberauer, K., & Lewandowsky, S. (2008). Forgetting in immediate serial recall: Decay, temporal distinctiveness, or interference? *Psychological Review, 115*, 544-576.

Oeusoonthornwattana, O., & Shanks, D. R. (2010). I like what I know: Is recognition a non-compensatory determiner of consumer choice? *Judgment and Decision Making, 5*, 310-325.

Oppenheimer, D. M. (2003). Not so fast! (and not so frugal!): Rethinking the recognition heuristic. *Cognition, 90*, B1-B9.

Pachur, T. (2010). Recognition-based inference: When is less more in the real world? *Psychonomic Bulletin & Review, 17*, 589-598.

Pachur, T. (2011). The limited value of precise tests of the recognition heuristic. *Judgment and Decision Making, 6*, 413-422.

Pachur, T., & Biele, G. (2007). Forecasting from ignorance: The use and usefulness of recognition in lay predictions of sports events. *Acta Psychologica, 125*, 99-116.

Pachur, T., Bröder, A., & Marewski, J. N. (2008). The recognition heuristic in memory based inference: Is recognition a non compensatory cue? *Journal of Behavioral Decision Making, 21*, 183-210.

Pachur, T., & Hertwig, R. (2006). On the psychology of the recognition heuristic: Retrieval primacy as a key determinant of its use. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*, 983-1002.

Pachur, T., Mata, R., & Schooler, L. J. (2009). Cognitive aging and the adaptive use of recognition in decision making. *Psychology and Aging, 24*, 901-915.

Pachur, T., Todd, P. M., Gigerenzer, G., Schooler, L. J., & Goldstein, D. G. (2011). The recognition heuristic: A review of theory and tests. *Frontiers in Cognitive Science, 2*, 1-14.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 534-552.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The Adaptive Decision Maker*. New York, NY: Cambridge University Press.

Pitt, M., Myung, I., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review, 109*, 472-491.

Pleskac, T. J. (2007). A signal detection analysis of the recognition heuristic. *Psychonomic Bulletin & Review, 14*, 379-391.

Pohl, R. F. (2006). Empirical tests of the recognition heuristic. *Journal of Behavioral Decision Making, 19*, 251-271.

Pohl, R. F. (2011). On the use of recognition in inferential decision making: An overview of the debate. *Judgment and Decision Making, 6*, 423-438.

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review, 111*, 333-367.

Reimer, T., & Katsikopoulos, K. V. (2004). The use of recognition in group decision-making. *Cognitive Science, 28*, 1009-1029.

Richter, T., & Späth, P. (2006). Recognition is used as one cue among others in judgment and decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*, 150-162.

Rieskamp, J., & Hoffrage, U. (1999). When do people use simple heuristics, and how can we tell? In G. Gigerenzer, P. M. Todd, & the ABC Research Group (Eds.), *Simple Heuristics that Make us Smart* (pp. 141-167). New York, NY: Oxford University Press.

Rieskamp, J., & Hoffrage, U. (2008). Inferences under time pressure: How opportunity costs affect strategy selection. *Acta Psychologica, 127*, 258-276.

Rieskamp, J., & Otto, P. E. (2006). SSL: A theory of how people learn to select strategies. *Journal of Experimental Psychology: General, 135*, 207-236.

Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. (2007). Cognitive Tutor: Applied research in mathematics education. *Psychonomic Bulletin & Review, 14*, 249-255.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review, 107*, 358-367.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*, 225-237.

Rumelhart, D. E., McClelland, J. L., & the PDP Research Group. (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition.* (Vol. 1). Cambridge, MA: MIT Press.

Salvucci, D. D. (2006). Modeling driver behavior in a cognitive architecture. *Human Factors, 48*, 362-380.

Salvucci, D. D., & Taatgen, N. A. (2008). Threaded cognition: An integrated theory of concurrent multitasking. *Psychological Review, 115*, 101-130.

Scheibehenne, B., & Bröder, A. (2007). Predicting Wimbledon 2005 tennis results by mere player name recognition. *International Journal of Forecasting, 23*, 415-426.

Schooler, L., & Hertwig, R. (2005). How forgetting aids heuristic inference. *Psychological Review, 112*, 610-628.

Schulte-Mecklenbeck, M., Kühberger, A., & Ranyard, R. (2010). *A Handbook of Process Tracing Methods for Decision Research*. New York, NY: Taylor & Francis.

Sedlmeier, P., & Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General, 130*, 380-400.

Sprenger, A. M. (2007). *Sequential Hypothesis Generation.* Doctoral dissertation, University of Maryland, College Park. Available from ProQuest Digital Dissertations database (UMI AAT 3260299).

Sprenger, A. M., & Dougherty, M. R. P. (2006). Differences between probability and frequency judgments: The role of individual differences in working memory capacity. *Organizational Behavior and Human Decision Processes, 99*, 202-211.

Taatgen, N. A., & Anderson, J. R. (2002). Why do children learn to say "broke"? A model of learning the past tense without feedback. *Cognition, 86*, 123-155.

Thagard, P. (1989a). Explanatory coherence. *Behavioral and Brain Sciences, 12*, 435-467.

Thagard, P. (1989b). Extending explanatory coherence. *Behavioral and Brain Sciences, 12*, 490-502.

Thagard, P. (2000). Probabilistic networks and explanatory coherence. *Cognitive Science Quarterly, 1*, 93-116.

Thagard, P., Kunda, Z., Read, S. J., & Miller, L. C. (1998). Making sense of people: Coherence mechanisms. In S. J. Read and L.C. Miller (Eds.), *Connectionist models of social reasoning and social behavior.* (pp. 3-26). Hillsdale, NJ: Erlbaum.

Thagard, P., & Shelley, C. P. (1997). Abductive reasoning: Logic, visual thinking, and coherence. In M.-L. Dalla Chiara (Ed.), *Logic and scientific methods* (pp. 413-427). Dordrecht: Kluwer.

Thomas, R. P., Dougherty, M. R. P., Sprenger, A. M., & Harbison, J. I. (2008). Diagnostic hypothesis generation and human judgment. *Psychological Review, 115*, 155-185.

Tomlinson, T., Marewski, J. N., & Dougherty, M. R. P. (2011). Four challenges for cognitive research on the recognition heuristic and a call for a research strategy

shift. *Judgment and Decision Making, 6*, 89-99.

Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological Review, 79*, 281-299.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5*, 207-232.

Vallesi, A., Shallice, T., & Walsh, V. (2007). Role of the prefrontal cortex in the foreperiod effect: TMS evidence for dual mechanisms in temporal preparation. *Cerebral Cortex, 17*, 466-474.

Van Maanen, L., & Marewski, J. N. (2009). Recommender systems for literature selection: A competition between decision making and memory models. In N. A. Taatgen & H. Van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 2914–2919). Austin, TX: Cognitive Science Society.

Volz, K. G., Schooler, L. J., Schubotz, R. I., Raab, M., Gigerenzer, G., & Von Cramon, D. Y. (2006). Why you think Milan is larger than Modena: Neural correlates of the recognition heuristic. *Journal of Cognitive Neuroscience, 18*, 1924-1936.

von Helversen, B., & Rieskamp, J. (2008). The mapping model: A cognitive theory of quantitative estimation. *Journal of Experimental Psychology: General, 137*, 73-96.

Wang, H., Johnson, T. R., & Zhang, J. (2006a). A hybrid system of abductive tactical decision making. *International Journal of Hybrid Intelligent Systems, 3*, 23-33.

Wang, H., Johnson, T. R., & Zhang, J. (2006b). The order effect in human abductive reasoning: An empirical and computational study. *Journal of Experimental & Theoretical Artificial Intelligence, 18*, 215-247.

Weber, E., Böckenholt, U., Hilton, D., & Wallace, B. (1993). Determinants of diagnostic hypothesis generation: Effects of information, base rates, and experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*, 1151-1164.